Implementació d'un sistema de crawlers per recollir dades en distints formats

Treball de final de grau

26/04/2017

Autor: **Gabriel Pieras Morell**

Director: **Jose María Barceló Ordinas**

Grau en Enginyeria Informàtica

Tecnologies de la Informació

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Agraïments

Aquest projecte no hauria estat possible sense l'inestimable ajuda dels professors del grup d'investigació, el Jose María Barceló i el Jorge García, que m'han guiat i impulsat per arribar fins al final.

També vull agrair la seva col·laboració a tots els meus companys de laboratori: Albert, Santiago, Pau, Aina i Gerard. Ells han ajudat a crear un gran ambient de companyonia i amistat a la universitat.

Per últim, a la meva família i amics, que m'han animat en tot moment i sense qui no hauria arribat on sóc.

Resum

Aquest treball intenta ampliar els serveis d'airACT i CommSensum automatitzant la recollida de dades de més zones de les disponibles. Per fer-ho, també s'haurà de muntar tot un sistema de servidors per al grup.

Resumen

Este trabajo intenta ampliar los servicios de airACT y CommSensum automatizando la recogida de datos de más zonas de las disponibles. Para hacerlo, también se tendrá que montar todo un sistema de servidores para el grupo.

Abstract

This project tries to extend the services provided by airACT and CommSensum, automatizing the data downloading service to act on more regions. To do this, it will also have to build a server for the group.

Índex

Contents

Índex 4
Índex de figures 7
Índex de taules 7
No table of figures entries found. .....**Error! Bookmark not defined.**
1. Introducció 8
1.1 Formulació del problema 8
1.2 Objectius 9
1.2.1 Objectius generals 9
1.2.2 Objectius tècnics 9
1.3 Competències tècniques 10
2. Gestió del projecte 10
2.1 Estat de l'art 10
2.1.1 Context 11
2.1.2 Actors implicats 11
2.1.3 Solucions existents 12
2.1.4 Tecnologies seleccionades 20
2.2 Abast 20
2.2.1 Abast del projecte 21
2.2.2 Possibles obstacles 21
2.3 Metodologia i rigor 22
2.3.1 Metodologia 22
2.3.2 Validació i seguiment 22
2.3.3 Eines de treball 23
2.4 Planificació 23
2.4.1 Definició de les tasques 23
2.4.2 Valoració d'alternatives i pla d'acció 27
2.4.3 Recursos 27
2.5 Pressupost 27
2.5.1 Estimació inicial del pressupost 28
2.5.2 Pressupost final 28
2.5.3 Control de gestió 30

3. Sostenibilitat i compromís social 30
3.1 Sostenibilitat ambiental 30
3.2 Sostenibilitat econòmica 31
3.3 Sostenibilitat social 31
3.4 Matriu de sostenibilitat 32
4. Tecnologies implicades 32
4.1 Python 33
4.X Javascript 33
4.2 Hipervisor 33
4.3 Citrix XenServer 34
4.4 Ubuntu 34
4.5 SSH 34
4.6 PostgreSQL 35
4.7 XML 35
4.8 CSV 35
4.9 HTTP 35
4.10 Node.js 35
5. Desenvolupament 36
5.1 Adquisició d'un nou servidor 36
5.1.1 Comparativa d'ofertes 37
5.1.2 Instal·lació de l'hipervisor i creació de màquines virtuals 38
5.2 Desenvolupament del *crawler* 41
5.2.1 Obtenció de les dades 41
5.2.2 Desenvolupament del *crawler* 43
5.3 Construcció del servidor intermediari de dades 46
5.3.1 Aïllament dels nodes 46
5.3.2 Enviament de les dades 47
6. Possibles millores 48
6.1 Inclusió de nous orígens de dades 48
6.2 Canvi de model de base de dades 48
6.3 Monitorització dels nodes de CAPTOR 48
7. Conclusions 49
Annex I. Llicències de dades 49

Annex II. Desplegament del *crawler* 50
Annex III. l2chroot 51
Bibliografia 52

Índex de figures

Figura 1. Estació automàtica de Palau Reial, situada al CSIC 8
Figura 2. Vista de la pàgina web de la Generalitat de Catalunya 12
Figura 3. Aplicació AireCat 13
Figura 4. Vista de la pàgina web de la comunitat de Madrid 13
Figura 5. Vista de la pàgina web de la Regió de Múrcia 14
Figura 6. Visor de la calidad del aire 15
Figura 7. Sistema CALIOPE 15
Figura 8 Pàgina de la llombardia 16
Figura 9. Pàgina de la EEA 17
Figura 10. Pàgina de WAQI 17
Figura 11. Pàgina d'air Quality Now 18
Figura 12. PlumeLabs 19
Figura 13. Funcionament de CommSensum 20
Figura 14. Diagrama de Gantt 26
Figura 15. Diferències entre hipervisors 34
Figura 16. Funcionament del projecte 36
Figura 17. Panell de control d'OVH 38
Figura 18. XenCenter 39
Figura 19. Creació d'una màquina virtual 40
Figura 20. MAC virtual del la màquina 40

Índex de taules

Taula 1. Hores per tasca 24
Taula 2. Calendari de tasques 25
Taula 3. Estimació inicial 28
Taula 4. Recursos humans 28
Taula 5. Recursos humans (2) 28
Taula 6. Recursos hardware 29
Taula 7. Recursos software 29
Taula 8. Costs indirectes 29
Taula 9. Cost total 30
Taula 10. Sobrecosts 30
Taula 11. Matriu de sostenibilitat 32
Taula 12. Comparativa d'ofertes 37
Taula 13. Paràmetres de la EEA 44

1. Introducció

1.1 Formulació del problema

La contaminació és un problema que afecta el planeta, però a la vegada té un impacte directe sobre la salut de les persones. En concret, la contaminació atmosfèrica provoca problemas respiratoris, malalties cardiovasculars, càncer (predominantment de pulmó) i, en general, una reducció de l'esperança de vida.

És per això que diverses organitzacions han posat en marxa distints operatius per tal de mesurar la pol·lució. En el cas de la Generalitat de Catalunya, la Xarxa de Vigilància i Previsió de la Contaminació Atmofèrica (XVPCA) ha col·locat una sèrie de sensors que mesuren, entre d'altres, les quantitats d'ozó troposfèric (O_3), de diòxid de nitrogen (NO_2), de monòxid de carboni i de partícules microscòpiques en suspensió.

Figura 1. Estació automàtica de Palau Reial, situada al CSIC

El gran problema d'aquestes estacions automàtiques és el seu preu. El pressupost de la XVPCA el 2014 va ser de 1'5M€[1], i el preu d'una sola estació automàtica pot ser de l'ordre dels centenars de milers d'euros[2].

És en aquest context on neix el projecte CAPTOR, un projecte europeu de col·laboració entre diverses entitats[3] que busca complementar aquests sensors amb uns de més barats, assequibles pels ciutadans, amb l'objectiu de formar una xarxa oberta de dades atmosfèriques on hi pugui accedir tothom[4]. CAPTOR està integrat amb la plataforma

CommSensum, una plataforma oberta de linked data on es publicaran les dades recollides[5,6].

El major problema dels sensors del projecte CAPTOR és la fiabilitat de les dades. No es pot assegurar d'entrada que uns sensors comercials de baix cost puguin recollir dades amb la mateixa precisió que els sensors de les estacions de referència, molt més cars i sofisticats. Per tal de corregir aquest problema es vol dotar CAPTOR amb la informació necessària per a poder comprovar si els sensors estan ben calibrats o si funcionen correctament.

Per altra banda, la infraestructura de software de què es disposa està sobredimensionada i ja no s'adapta a les necessitats del grup d'investigació. Es perden temps i diners formant els becaris nous que entren al grup, ja que la plataforma està molt mal documentada, i cada vegada que es vol afegir alguna funcionalitat s'han d'arreglar tots els problemes que sorgeixen sense saber exactament què ha fallat i per on començar a arreglar-ho.

Per últim, el grup només compta amb un únic servidor, una màquina virtual del departament d'arquitectura de computadors, que no té la capacitat de processament, la memòria o l'emmagatzematge suficients com per allotjar els nous serveis, a més d'un període de manteniment a l'estiu, una etapa crítica del projecte.

1.2 Objectius

1.2.1 Objectius generals

L'objectiu principal d'aquest projecte és conscienciar a la població dels problemes que comporta la contaminació atmosfèrica, fent visible la contaminació a la que estem sotmesos cada dia. Per a això alimentarem amb noves dades la web http://www.airact.org, una web desenvolupada dins el grup per un altre estudiant i impulsada per Ecologistas en Acción que mostra les dades de contaminació atmosfèrica de Catalunya i Madrid.

Aquestes noves dades beneficiaran a totes aquelles persones que lluiten per informar i conscienciar dels problemes que patim a causa de la pol·lució, i facilitaran la difusió d'aquesta informació a la resta de la població.

1.2.2 Objectius tècnics

En primer lloc es vol ampliar el servei de recollida de dades que té el grup, que actualment recull les dades de Catalunya i Madrid que s'han mencionat anteriorment. Per fer-ho es desenvoluparà un sistema automàtic que es descarregui les dades de contaminació, horàries i en temps real, dels diferents països involucrats en el projecte: Àustria, Espanya i Itàlia. Aquestes dades s'aniran desant en una base de dades PostgreSQL i s'enviaran, a través del servei web de CommSensum, a un base de dades que compartiran amb els nodes de CAPTOR. D'aquesta manera es podran comparar

fàcilment les dades de les estacions calibrades i mantingudes per professionals amb les dels sensors més assequibles i que cadascú podrà tenir a casa seva.

En segon lloc, es construïrà un servei que permeti els nodes de CAPTOR emmagatzemar les dades en format CSV en un servidor propi. A aquest servidor hi tendran accés les persones responsables de fer la calibració. Aquest servidor també s'encarregarà d'enviar les dades que rebi a la plataforma CommSensum.

Per tal de portar a terme aquest objectiu també serà necessari aconseguir un hardware més potent que el que hi ha actualment, que com ja s'ha mencionat no és suficient. Aquest hardware servirà per allotjar-hi els serveis ja mencionats, així com d'altres de què disposa el grup.

Per últim, una vegada adquirit el hardware, es volen posar en marxa diferents màquines virtuals pels diferents serveis de què disposa el grup, així com aconseguir una bona inegració dels serveis que treballin junts.

1.3 Competències tècniques

A continuació es justifiquen les competències tècniques del projecte.

Per l'adquisició d'un servidor adequat a les necessitats del grup, així com del muntatge de les màquines virtuals:

- **CTI1.3:** Seleccionar, desplegar, integrar i gestionar sistemes d'informació que satisfacin les necessitats de l'organització amb els criteris de cost i qualitat identificats. [Bastant]

Per el desenvolupament del crawler:

- **CTI3.1:** Concebre sistemes, aplicacions i serveis basats en tecnologies de xarxa, tenint en compte Internet, web, comerç electrònic, multimèdia, serveis interactius i computació ubiqua. [En profunditat]

Per el muntatge del servidor de CSV i la gestió de permisos i usuaris que s'hi fa:

- **CTI4:** Emprar metodologies centrades en l'usuari i l'organització per al desenvolupament, l'avaluació i la gestió d'aplicacions i sistemes basats en tecnologies de la informació que assegurin l'accessibilitat, l'ergonomia i la usabilitat dels sistemes. [Una mica]

2. Gestió del projecte

2.1 Estat de l'art

En aquesta secció es descriuen el context del projecte, les solucions tecnològiques ja existents i les tecnologies que finalment s'han seleccionat.

2.1.1 Context

La legislació vigent de la Unió Europea estableix que les autoritats nacionals són responsables d'avaluar la qualitat de l'aire i d'informar-ne a la població. Alhora, la llei espanyola delega aquestes tasques a les administracions de les comunitats autònomes, sempre dins l'àmbit de les seves competències. És per això que els diferents governs han posat en marxa sistemes de control de la qualitat de l'aire i tambe pàgines web que informen de les mesures que prenen.

Aquesta llei, però, provoca que la informació de primera mà estigui molt fragmentada i hem d'acudir a pàgines agregadores, on podem accedir al conjunt de la informació d'una manera més senzilla.

Per altra banda, hi ha iniciatives privades que també recullen informació sobre la pol·lució, ja sigui amb la informació oficial o amb mesures pròpies, i la mostren en pàgines web.

Aquesta informació pública l'utilitzen els grups ecologistes per dur a terme campanyes de conscienciació i accions reivindicatives, si bé és difícil arribar a la població que no està interessada en aquesta problemàtica en primer lloc. Així mateix, els darrers mesos ha estat notícia que alguns ajuntaments han restringit la circulació de vehicles dins de les ciutats a causa de la contaminació. No obstant, els casos són aïllats i la població ho percep més com una molèstia que com una mesura necessària.

Així doncs, és important que existeixi una eina de fàcil ús i accés i que disposi de tota la informació rellevant.

2.1.2 Actors implicats

El projecte està dirigit, en primer lloc, al grup de recerca SANS (Statistical Analysis of Networks and Systems) del departament d'Arquitectura de Computadors de la UPC. Ells seran els que utilitzaran les dades de les zones on s'hi desplegui el projecte CAPTOR per fer la calibració. També serà el grup d'investigació el que s'aprofitarà de la renovació i expansió del *back-end*.

En segon lloc, les organitzacions ecologistes que col·laboren en el projecte (Ecologistas en Acción, Legambiente Onlus i Global-2000 Umweltschutzorganisation) tenen interès en saber que els sensors funcionen correctament, ja que amb la informació que obtinguin del projecte poden pressionar els responsables de la qualitat de l'aire per tal de que prenguin mesures.

En tercer lloc, els usuaris finals dels sensors de CAPTOR i d'airACT seran ciutadans corrents dels que no s'espera que tinguin coneixements tècnics. Aquests ciutadans seran els que compraran els sensors i col·laboraran a crear la xarxa oberta de dades atmosfèriques.

Per últim, el rol de desenvolupador serà dut a terme per una sola persona, tot i que el correcte compliment dels terminis determinarà el correcte funcionament del projecte sencer.

2.1.3 Solucions existents

Aquest projecte no és en cap cas una idea novedosa. De fet, existeixen diverses solucions que ofereixen dades de contaminació atmosfèrica. En aquest apartat es presenten amb una mica de detall les solucions més rellevants a les zones que ens interessen.

- **Pàgines web de les comunitats autònomes**

  Les pàgines pròpies de les comunitats autònomes mostren, amb distints nivell de detall, les dades que recullen amb les seves estacions. A continuació se'n mostren només tres exemples:

  El web de la Generalitat de Catalunya proporciona accés a les dades recollides a les estacions de què disposa la Xarxa de Vigilància i Previsió de la Qualitat de l'Aire. S'hi mostren les dades numèriques en temps real i permet descarregar-se un històric de dades, però la pàgina és una mica lenta a l'hora de respondre als clics de l'usuari. A més, accedir-hi des de la pàgina web de la Generalitat no és senzill, ja que es troba bastant amagada i s'ha de navegar força per arribar-hi si no en coneixem l'adreça.

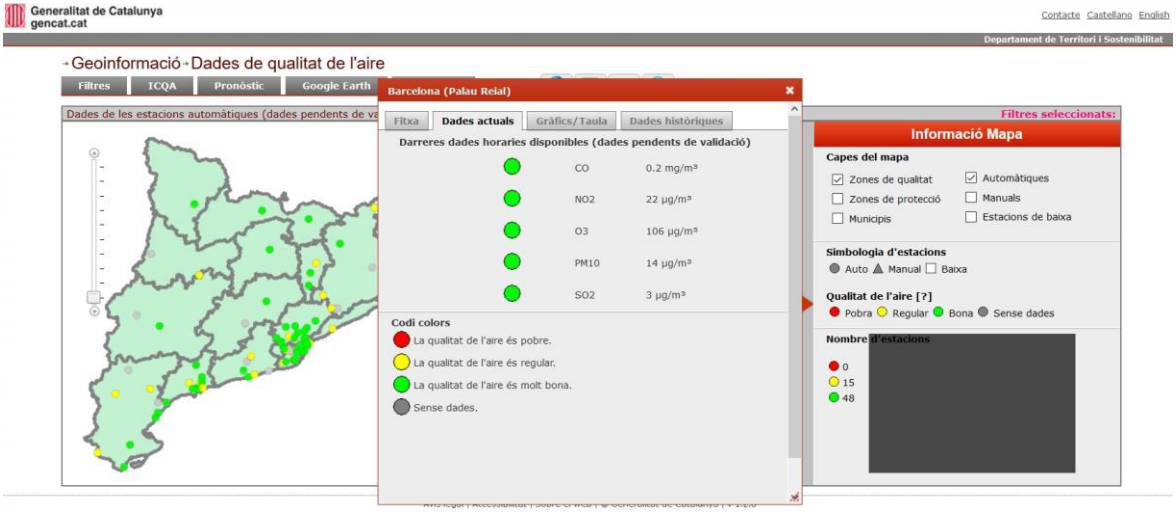


Figura 2. Vista de la pàgina web de la Generalitat de Catalunya

La Generalitat també disposa d'Aire.cat, una aplicació mòbil per Android i iOS que mostra les mateixes dades.

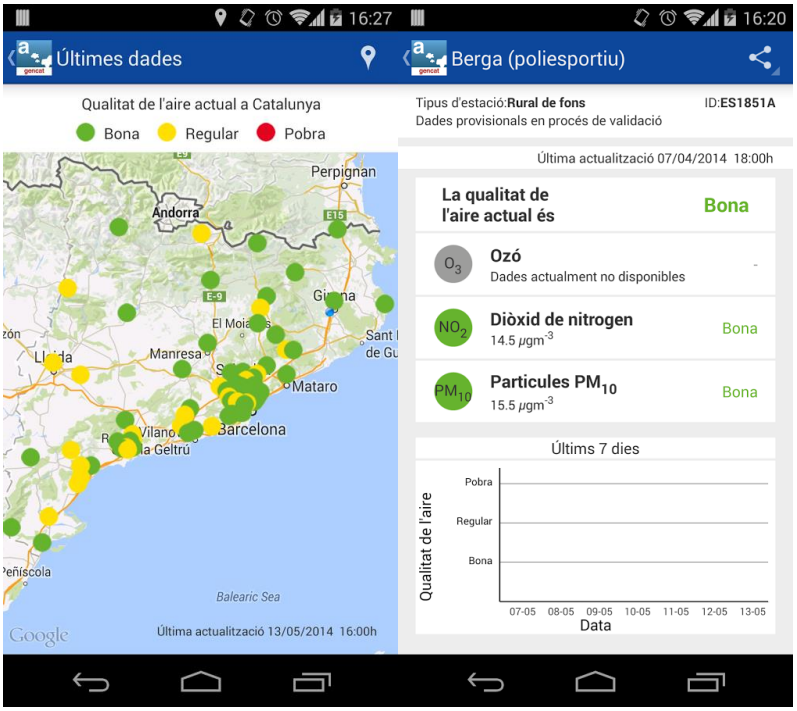


Figura 3. Aplicació AireCat

La comunitat de Madrid també disposa de la seva pàgina d'informació, on hi mostra les dades numèriques i també gràfiques amb l'evolució temporal dels contaminants. Aquestes dades, però, es mostrenen forma d'imatge, de manera que només són llegibles per humans. L'ajuntament de Madrid també disposa d'una aplicació mòbil, però les valoracions són generalment dolentes.

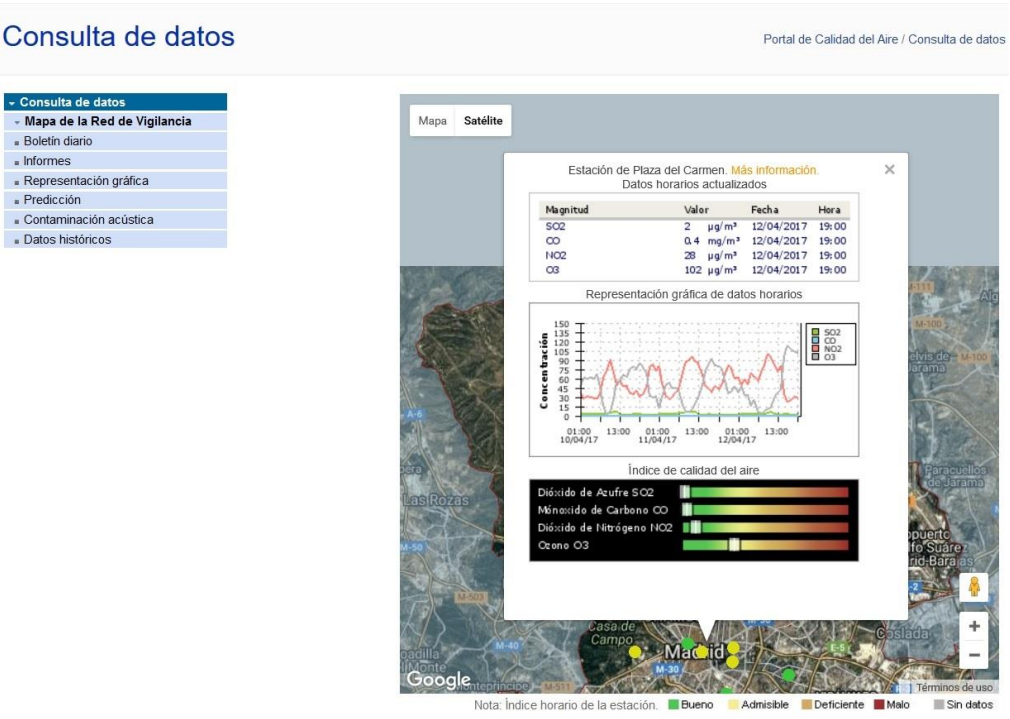


Figura 4. Vista de la pàgina web de la comunitat de Madrid

La pàgina web de la qualitat de l'aire a Múrcia també mostra les dades que recullen sobre un mapa de la regió. Igual que en el cas anterior, les dades només són llegibles per humans, i per obtenir-ne les dades automàticament és necessari fer-ho a través d'una eina poc intuïtiva.

Figura 5. Vista de la pàgina web de la Regió de Múrcia

- **Visor de la calidad del aire (Ministeri de medi ambient)**

  Les dades de les comunitats autònomes s'envien a un servei del Ministeri d'agricultura i pesca, alimentació i medi ambient. Aquest visor mostra les dades que recopila sobre un mapa d'Espanya, però accedir a les dades numèriques torna a ser lent i poc intuïtiu.

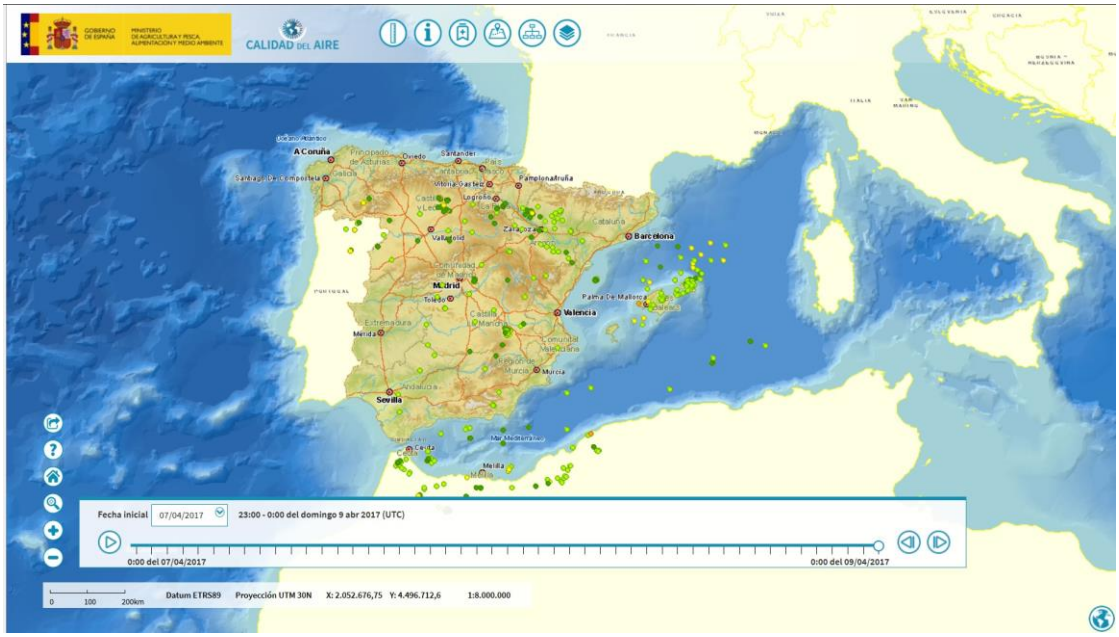


Figura 6. Visor de la calidad del aire

- **Sistema CALIOPE**

El sistema CALIOPE és un projecte del Ministeri d'agricultura i pesca, alimentació i medi ambient amb el *Barcelona Supercomputing Center* (BSC), que no té com a objectiu pricipal mostrar la contaminació actual, sinó mostrar el pronòstic de la qualitat de l'aire.

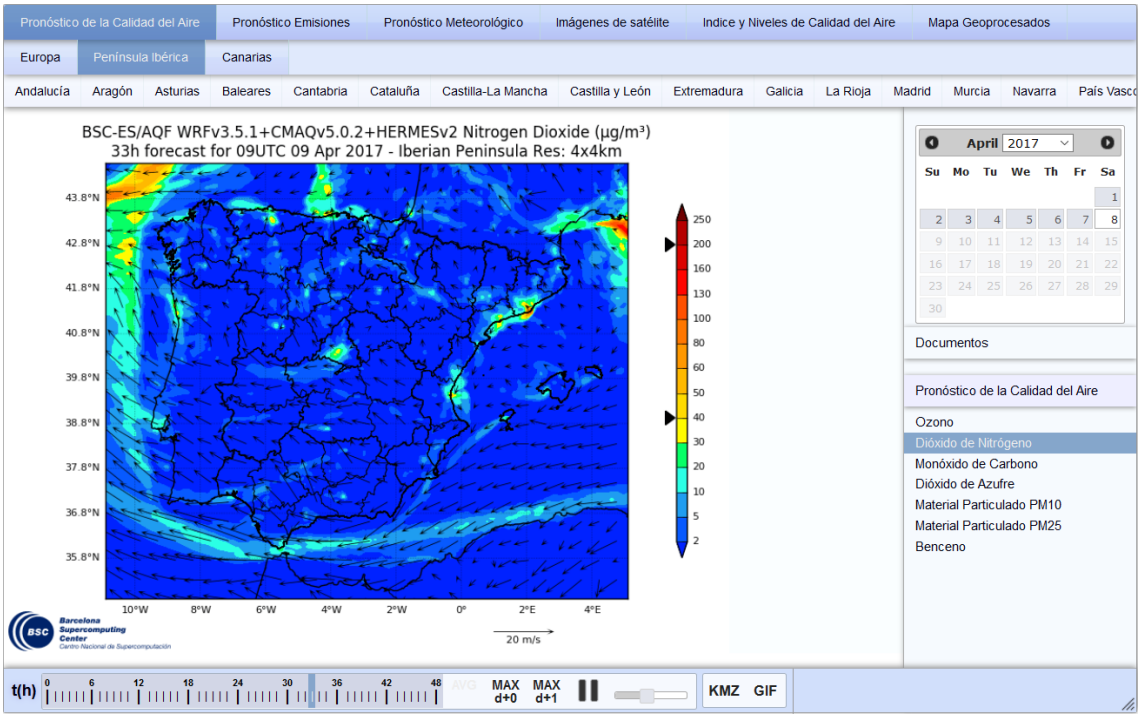


Figura 7. Sistema CALIOPE

- **Pàgines de l'administració italiana**

  A Itàlia, una altra zona rellevant per al projecte CAPTOR, l'única zona que recull dades horàries en temps real de contaminació és la regió de la Llombardia. Malauradament, disposa d'un sistema contra *bots* que impedeix la descàrrega de dades de forma automàtica.

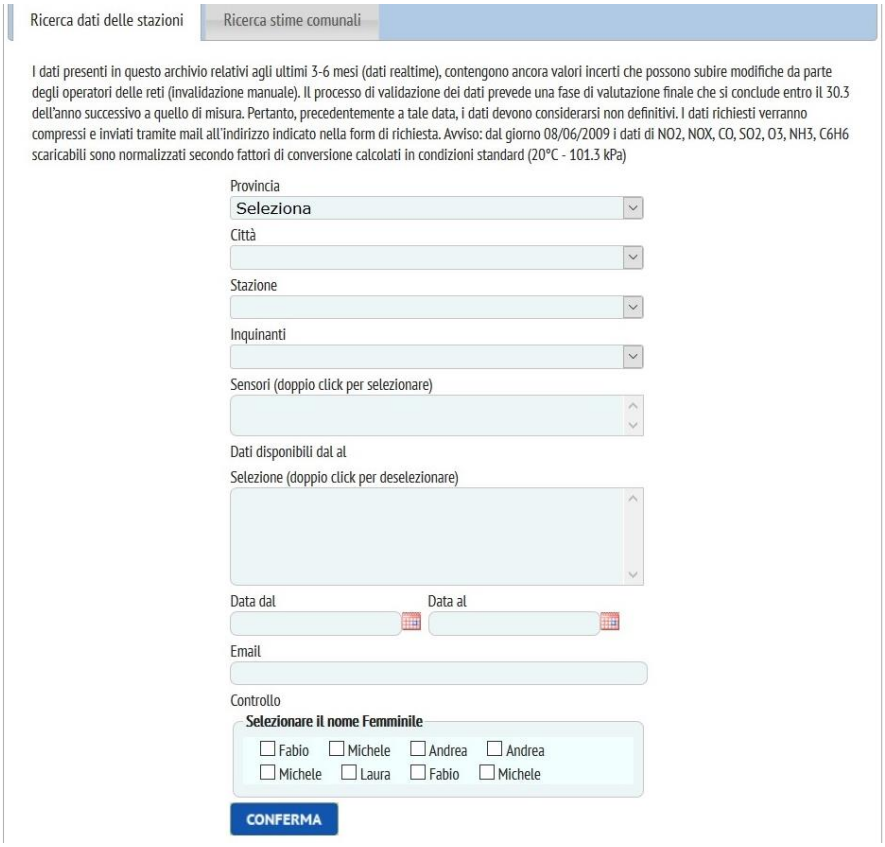


Figura 8 Pàgina de la llombardia

- **Servei de la European Environment Agency**

  Englobant tots aquests serveis la *European Environment Agency* (EEA) també disposa d'una base de dades on recullen tota la informació disponible dels països membres de la UE. Aquestes dades estan disponibles a través d'una API molt senzilla i fins el març de 2017 es mostraven en un mapa com en els exemples anteriors.

Download of UTD Air Quality data

This form provides the interface for accessing the download service provided by EEA in order to download UTD air quality data. For details on this download service, please see UTDAirQualityDownloadGuide.pdf

Note: Please notice the new option to make asynchronous requests. See the guideline.

FromDate (yyyy-mm-dd hh:mm):	2014-11-10
ToDate (yyyy-mm-dd hh:mm):	2014-11-11
Countrycode (comma separated for multiple):	dk
InsertedSinceDate (yyyy-mm-dd):	
UpdatedSinceDate (yyyy-mm-dd):	
Pollutant (comma separated for multiple):	
Namespace (only one):	
Format (XML, CSV, CSV_OzoneWeb, CSV2):	CSV
Qoute_output_values: (Yes, No):	No
RunAsync: (True, False):	False
UserToken (mandatory):	
	Run

Figura 9. Pàgina de la EEA

- **World Air Quality Index**

El *World Quality Air Index* és un projecte nascut a la Xina que actualment conté dades de més de 70 països i més de 9000 estacions de referència. Aquesta pàgina ofereix les últimes dades disponibles de cada punt en un mapa i a través d'un servei de descàrrega de dades, tot i que no diferencia si són mitjanes horàries, octohoràries o diàries.

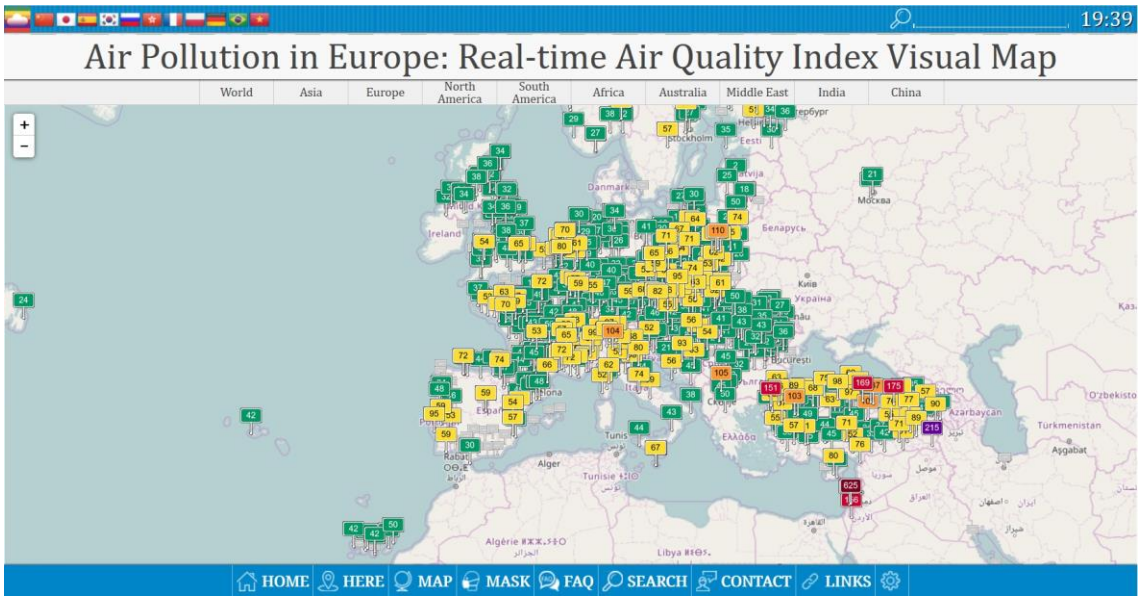


Figura 10. Pàgina de WAQI

- **Air Quality Now**

  Air Quality Now disposa principalment d'estacions a França, a més d'unes quantes a la resta d'Europa. Conviden a les ciutats a adherir-s'hi i enviar les seves dades, i no sembla que ho facin pel seu compte.

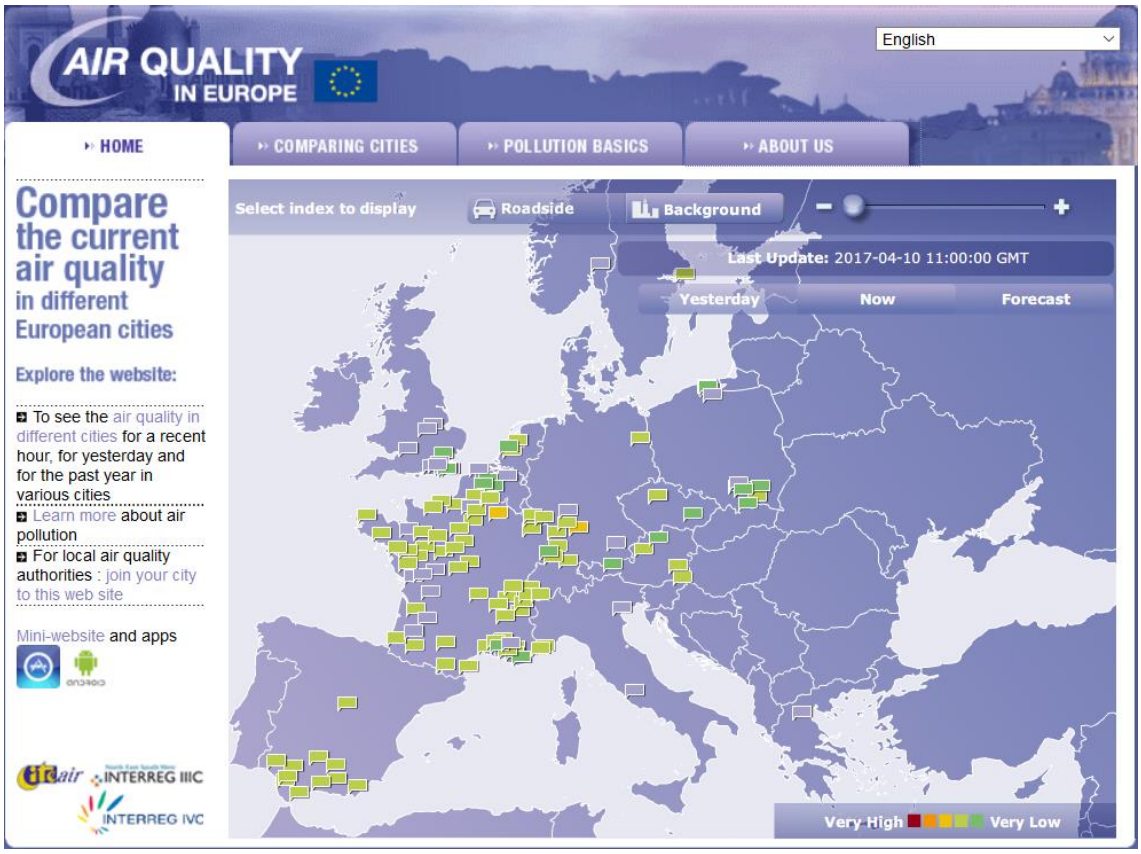


Figura 11. Pàgina d'air Quality Now

airqualitynow.eu

- **Plume Labs**

  Plume Labs és una iniciativa que recull dades d'arreu del món i les presenta segons un "Plume index", una escala pròpia que té una correspondència desconeguda amb les unitats del sistema internacional. A més, la decisió de representar les ciutats segons el número d'habitants resulta en un mapa confús i, si l'usuari no s'apropa, les ciutats petites queden ocultes sota les més grans.

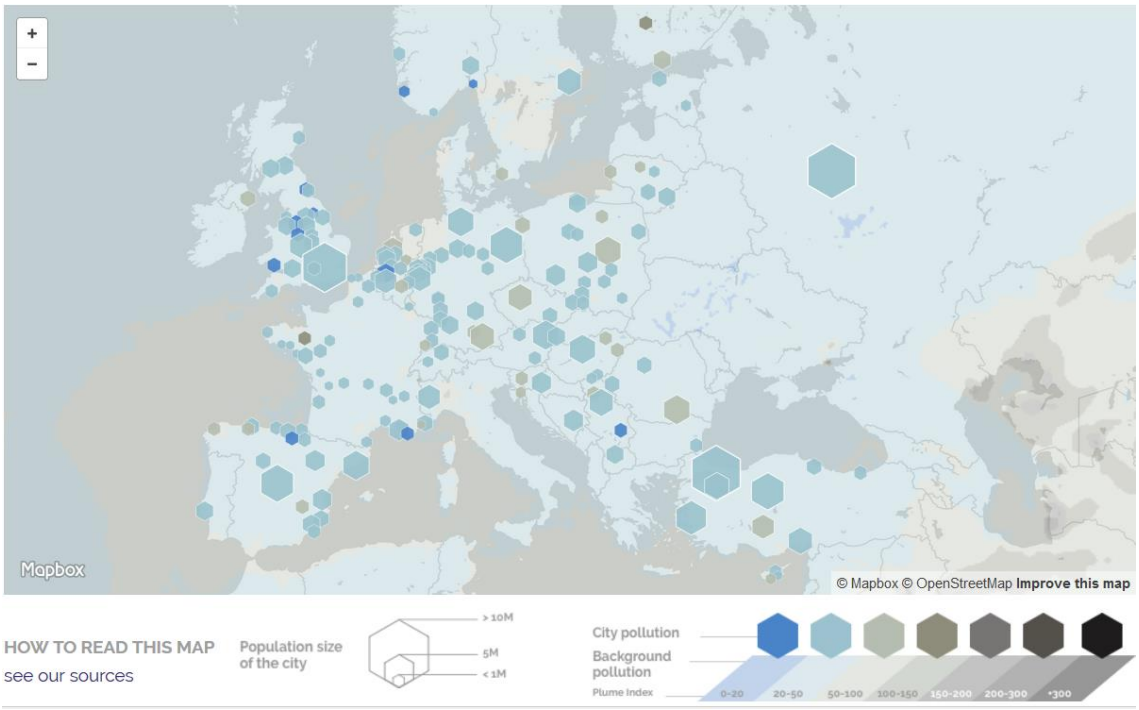


Figura 12. PlumeLabs

- **CommSensum / airACT**

  CommSensum és una plataforma *Linked Data* desenvolupada per la UPC que enfocada a emmagatzemar la informació que rep de nodes distribuïts, és a dir, *Internet of Things (IoT)*. Aquesta plataforma emmagatzema les dades en una base de dades mySQL i les publica en RDF. Disposa d'una API escrita en Node.js, un llenguatge que permet fer crides asíncrones i facilita l'enviament d'un gran número de peticions.

  A la figura següent es pot veure el seu funcionament de forma esquemàtica. En primer lloc, a l'esquerra hi apareixen les fonts de dades: els nodes distribuïts, com els del projecte CAPTOR; dades en distints formats provinents de distintes fonts, com poden els projectes mencionats anteriorment; o altres tipus de dades IoT, com les que puguin recollir els sensors d'un mòbil. Aquestes dades s'envien, a través de la API, a CommSensum, on s'emmagatzemen en un únic format. Per últim, les aplicacions web i mòbils obtenen les dades que necessiten a través d'altres peticions a la API i les mostren al públic.

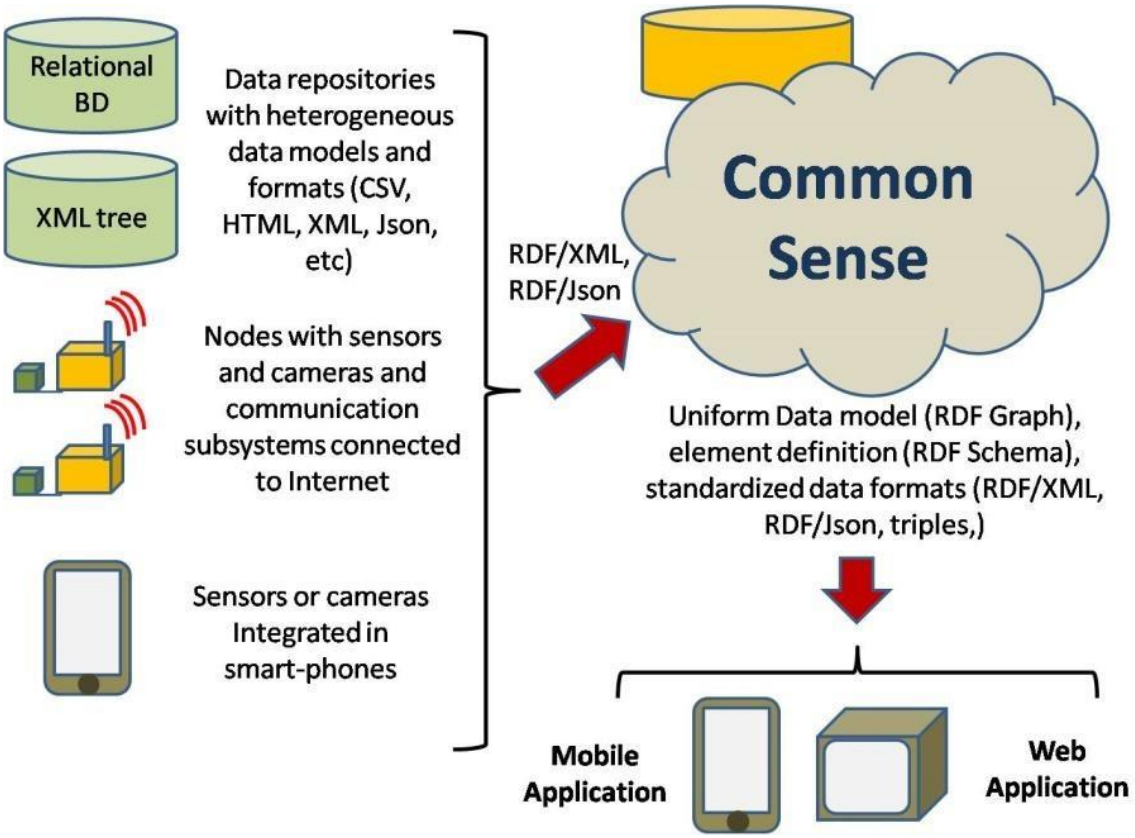


Figura 13. Funcionament de CommSensum

AirACT és una d'aquestes aplicacions que obtenen les dades de CommSensum. Desenvolupada a la UPC per a Ecologistas en Acción, compta amb una pàgina web i una aplicació d'Android, i mostra sobre un mapa les dades de contaminació atmosfèrica captades a les estacions de referència de Catalunya i Madrid, les úniques disponibles a CommSensum.

2.1.4 Tecnologies seleccionades

Una vegada comparades les opcions ja existents s'ha considerat que el més adeqüat és ampliar les dades d'estacions de referència amb què compta CommSensum, i mostrar-les a través d'airACT.

La font de les dades serà el servei de la EEA, que en permet la publicació sempre que es segueixi la seva política de dades, i que disposa de totes les dades horàries i en temps real que recullen les administracions públiques a Europa.

2.2 Abast

En un projecte gran és important determinar-ne l'abast. Això permet posar uns límits a l'ambició i permet assegurar-ne la finalització. Així mateix, és important identificar els possibles obstacles que puguin sorgir durant el projecte i saber com encarar-los.

2.2.1 Abast del projecte

Tenint en compte els objectius del projecte s'han determinat els següents punt que determinen l'abast del projecte.

- Adquirir un servidor per al grup, on s'hi muntaran màquines virtuals per als projectes del grup. Com que la compra d'un servidor per a cada projecte seria massa cara, una màquina més potent amb un hipervisor és una solució adeqüada.

- Obtenir les dades de qualitat de l'aire dels països implicats en el projecte CAPTOR. Les dades de les estacions de referència seran clau a l'hora de calibrar els sensors, de manera que el projecte es centrarà en obtenir les dades d'Espanya, Àustria i Itàlia. Es dissenyarà un script en python que reculli les dades i les desi de manera ordenada a la base de dades per, posteriorment, envar-les a CommSensum.

- Muntar un servidor on els nodes de CAPTOR hi puguin desar les dades en format CSV. Aquest format és fàcil de llegir en python i es pot importar a un full de càlcul, eina que encara empren algunes persones que treballen amb aquestes dades. Les dades quedaran desades en format CSV i també s'enviaran a CommSensum.

- Assegurar la bona integració entre CommSensum, airACT i els nodes de CAPTOR una vegada inserides les noves dades. No té sentit crear una nova aplicació que mostri les dades si ja n'existeix una que s'alimenta de les dades de CommSensum. Per això s'ampliarà la informació de la que s'alimenta airACT i es posarà especial atenció en la integració dels serveis.

2.2.2 Possibles obstacles

Aquest projecte depèn completament de les dades de tercers, de manera que la impossibilitat d'accedir a aquestes dades és el principal obstacle al que ens encarem. Si això ocorregués podria provocar un retard important en el calendari del projecte. Per tant, la millor solució seria posar-ho en coneixement de la resta d'integrants del grup, que podrien ajudar a cercar alternatives, o valorar l'opció de deixar la zona de banda per centrar-se en altres parts del projecte.

Un altre problema a tenir en compte és l'escalabilitat de l'aplicació. Actualment les dades que s'insereixen a CommSensum són poques i el hardware del servidor és molt limitat, i és possible que la plataforma no suporti la quantitat de dades noves que s'obtinguin tot i actualitzar el hardware. Una possible solució passaria per revisar el codi de la API i mirar d'optimitzar la inserció de dades.

Amb l'increment del volum de dades també cal plantejar-se si una base de dades relacional és la millor solució a llarg termini. A mesura que el volum de les dades creix el rendiment de la base de dades decau, i l'escalabilitat de l'aplicació es redueix. Una

solució és plantejar-se l'ús de bases de dades no relacionals, com poden ser mongoDB, Redis o Cassandra.

Per últim, i en menor mesura, un altre problema és el desconeixement de la tecnologia utilitzada. En aquest cas, la solució és dedicar hores a l'aprenentatge autònom.

2.3 Metodologia i rigor

En aquesta secció s'especifiquen la metodologia i eines de treball i quins mètodes de validació i seguiment s'han fet servir durant el projecte.

2.3.1 Metodologia

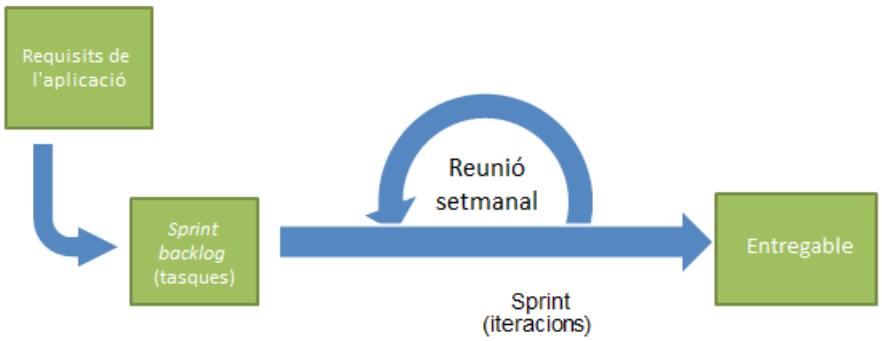
Triar una bona metodologia és vital per al bon desenvolupament del projecte. En aquest cas s'ha decidit seguir una metodologia àgil, que permet afrontar les modificacions del projecte més fàcilment. En concret s'ha adoptat una variació de *Scrum* que no tendrà reunions diàries.

S'han definit tres punts importants:

En primer lloc, dividir la feina en tasques, com més petites millor. Això es durà a terme amb l'ajuda dels professors responsables del grup, i evitarà que es creïn tasques innecessàries.

En segon lloc, ordenar les tasques segons l'urgència a partir dels terminis i les dependències entre tasques. Una vegada ordenades es posaran en un *backlog*.

Per últim, realitzar les tasques d'una en una. Si els requisits canvien durant l'execució del projecte és tan senzill com modificar les tasques pendents o afegir-ne de noves, sempre tenint-ne en compte la prioritat.



2.3.2 Validació i seguiment

Cada setmana es faran reunions amb els professors i la resta d'integrants del grup de recerca. En aquestes reunions cada integrant del grup informarà del seu progrés durant la setmana, així com dels possibles problemes que hagi pogut tenir. Això permet que els companys puguin aportar nous punts de vista als problemes i fer suggeriments per millorar el projecte.

Aquestes reunions són importants, perquè faciliten seguir i validar l'estat del projecte en tot moment. A més, permet que tots els integrants estiguin al corrent de l'estat dels altres projectes, ja que alguns d'ells estan relacionats i tenen parts comunes.

2.3.3 Eines de treball

Aquest projecte està integrat a dins del projecte CAPTOR i CommSensum i hi ha parts en comú amb altres integrants de l'equip. És per això que resulta necessari usar eines que facilitin la comunicació entre els membres i la gestió del projecte.

Per a la gestió de codi s'utilitzarà un repositori de git. En concret s'usarà **BitBucket**, que és el servei on hi ha el codi d'altres serveis del grup. El control de versions permetrà tornar ràpidament a una versió entiga del codi en cas d'error, i el fet de ser una eina en línia permet que el codi estigui disponible per fer-hi feina des de distints llocs.

Pel seguiment del projecte s'usarà **Trello**. Trello compta amb una aplicació web i aplicacions mòbils, i permet organitzar un *backlog* amb les tasques de manera molt intuïtiva.

També s'utilitzarà **Google Drive** per posar documents en comú entre els membres del grup. Drive, a més, permet veure i editar alguns formats sense necessitat de descarregar-los.

Per últim, s'utilitzarà **Slack** per mantenir la comunicació de manera informal entre els membres del grup. Slack és una eina gratuïta per a petits grups que permet tenir diverses sales de xat, i té aplicacions web, d'escriptori i mòbils. Per a comunicació mes formal s'utilitzarà el correu electrònic.

2.4 Planificació

Aquest projecte va començar el febrer del 2016 amb la intenció de presentar-lo el juny o l'octubre del mateix any, però problemes amb l'obtenció de les dades varen provocar que s'endarrerís. En aquesta secció es descriu la planificació final del projecte.

2.4.1 Definició de les tasques

A continuació s'expliquen les tasques en què s'ha dividit el projecte:

Tasca	Temps (h)
Familiarització amb les tecnologies i els llenguatges usats	80
Recerca	20
Aprenentatge autònom	60
Gestió de projectes	75
Anàlisi i documentació del codi existent	35
Desenvolupament del crawler	170
Adquisició de les dades	40
Disseny	40
Implementació	80
Testing i millores	10
Millores a la API de CommSensum	30
Adquisició d'un nou servidor	40
Recerca d'alternatives	20
Instal·lació de l'hipervisor i creació de màquines virtuals	20
Posada en marxa d'un servidor intermediari de dades	70

Creació d'usuaris	10
Creació de mesures d'aïllament i seguretat	30
Implementació d'un script d'enviament de dades	30
Documentació, redacció de la memòria i preparació de la defensa	100
Total	600

Taula 1. Hores per tasca

- **Familiarització amb les tecnologies i els llenguatges usats**
  És vital tenir un coneixement adequat de les tecnologies a utilitzar durant el projecte. Per això es van dedicar diverses hores a aprendre Python abans de començar el projecte.

- **Gestió de projectes**
  La gestió del projecte es durà a terme a través de l'assignatura de GEP. S'analitzaran els requisits del projecte i també els riscos que hi pugui haver. A més, s'elaborarà una planificació temporal i es faran estudis de costos i sostenibilitat.

- **Anàlisi i documentació del codi existent**
  CommSensum és una plataforma gran i no està pertinentment documentada. Serà necessari analitzar bé el codi per no cometre errors que puguin posar en perill la integritat de la plataforma.

- **Desenvoupament del crawler**
  La tasca principal i més llarga del projecte és desenvolupar el crawler. Aquesta tasca engloba el disseny i la implementació del crawler, però també l'adquisició de les dades, un punt crític del projecte. La seva execució depèn de les tasques anteriors.
  Precisament l'adquisició de les dades ha endarrerit el projecte uns quants mesos, ja que fins després de l'estiu de 2016 no es va aconseguir un origen fiable de dades horàries en temps real. A la taula 1 hi apareixen el número d'hores estimades de feina, no el temps d'espera. El període d'espera sí que apareix reflectit al diagrama de Gantt de la figura 14.

- **Millores a la API de CommSensum**
  Aquesta tasca conté totes les millores necessàries a la API per fer que el projecte funcioni correctament.

- **Adquisició d'un nou servidor**
  Una altra tasca és la compra del servidor. Es cercaran distintes opcions per trobar l'oferta que més s'adapti a les necessitats del grup. El major entrebanc en aquest pas serà la burocràcia de la UPC per realitzar els pagaments, però no s'hauria d'endarrerir més d'un o dos dies. Després es triarà quin hipervisor convé instal·lar i es configurarà la màquina.

- **Posada en marxa d'un servidor intermediari de dades**
  Una de les màquines virtuals s'usarà per a muntar-hi un servidor on els nodes de CAPTOR hi puguin enviar dades. Es crearan usuaris per a tots els nodes i s'implementaran mesures de seguretat per evitar que puguin interferir entre ells. Per acabar es farà un *script* per enviar aquestes dades a CommSensum. Aquesta tasca depèn de l'adquisició del servidor, ja que amb l'actual és impossible fer-ho.

- **Documentació, redacció de la memòria i preparació de la defensa**
  La documentació s'anirà escrivint amb el projecte, mentre que la memòria s'escriurà una vegada el projecte estigui acabat. La defensa, juntament amb la presentació en powerpoint, es prepararà al final. Aquest pas depèn de tots els anteriors per a poder realitzar-lo.

En total s'estima que el temps total que es dedicarà al projecte són unes 600 hores. Suposant una dedicació d'uns 4 hores diàries durant els dies hàbils podem calcular una durada d'aproximadament set mesos, si bé aquest període s'ha de repartir entre les tasques que es varen realitzar la primavera passada i les que es varen començar una vegada acabat l'estiu.

A continuació s'inclou una taula amb les dates aproximades d'inici i final de cada tasca i el diagrama de Gantt resultant d'aquesta planificació. Per a major claritat s'ha codificat cada tasca amb un color diferent.

	Familiarització amb les tecnologies i els llenguatges usats	**01/02/16**	**11/03/16**
	Recerca	01/02/16	18/02/16
	Aprenentatge autònom	19/02/16	11/03/16
	Gestió de projectes	**15/02/16**	**01/04/16**
	Anàlisi i documentació del codi existent	**20/03/16**	**07/04/16**
	Desenvolupament del sistema de crawlers	**07/04/16**	**21/12/16**
	Adquisició de les dades	07/04/16	13/10/16
	Disseny	07/11/16	22/11/16
	Implementació	23/11/16	16/12/16
	Testing i millores	17/12/16	21/12/16
	Adquisició d'un nou servidor	**15/09/16**	**22/09/16**
	Recerca d'alternatives	15/09/16	17/09/16
	Muntatge de màquines virtuals	17/09/16	22/09/16
	Posada en marxa d'un servidor intermediari de dades	**20/10/16**	**04/11/16**
	Creació d'usuaris	20/10/16	21/10/16
	Creació de mesures de seguretat i aïllament	24/10/16	28/10/16
	Creació d'script d'enviament de dades	31/10/16	04/11/16
	Millores API de CommSensum	**23/09/16**	**19/10/16**
	Correcció d'errors a l'API	23/09/16	29/09/16
	Implementació de noves funcions i millores a l'API	30/09/16	19/10/16
	Documentació i redacció de la memòria	**08/01/17**	**17/02/17**
	Redacció de la memòria	08/01/17	03/02/17
	Preparació de la defensa	06/02/17	17/02/17

Taula 2. Calendari de tasques

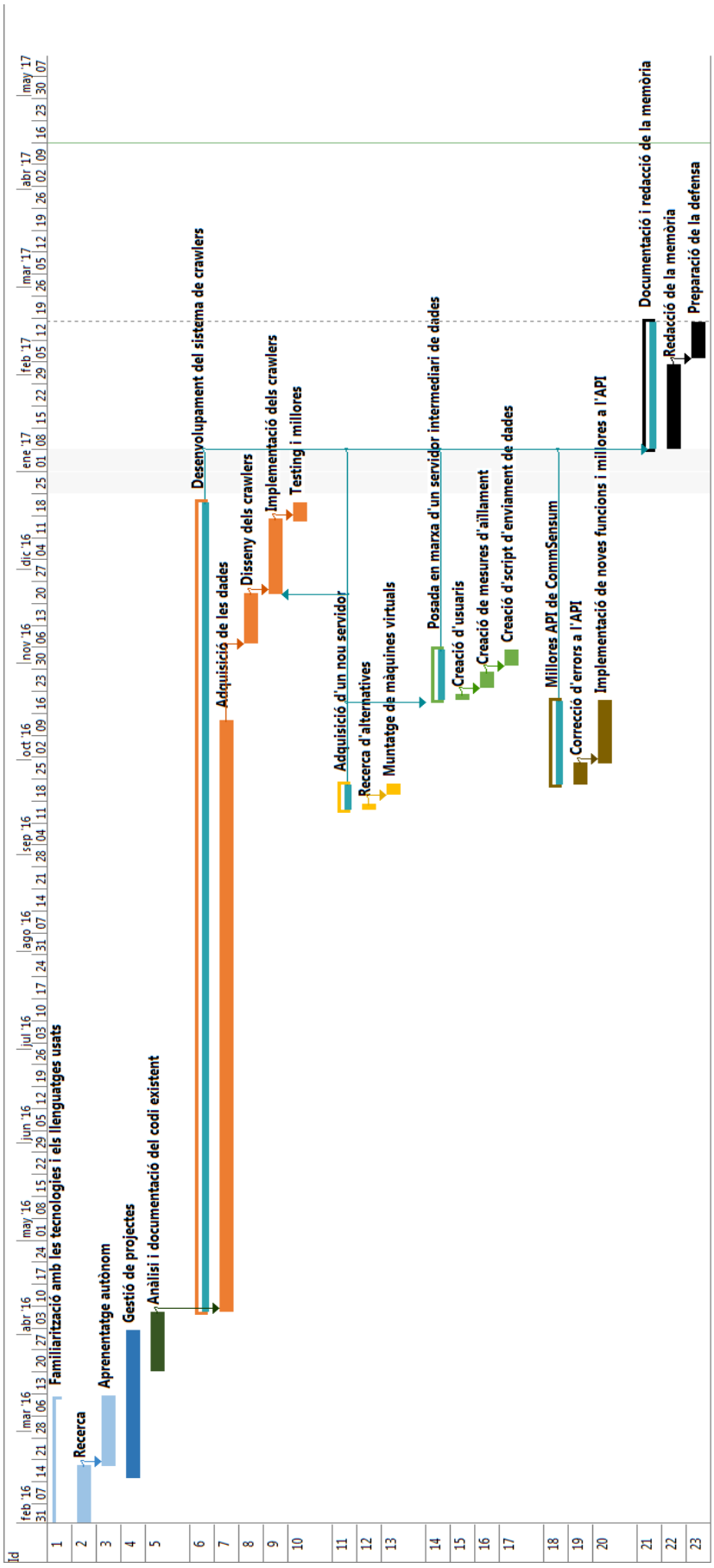


Figura 14. Diagrama de Gantt

2.4.2 Valoració d'alternatives i pla d'acció

Els obstacles que puguin aparèixer durant l'execució del projecte poden introduir desviacions en el compliment del calendari.

Per una banda, el desconeixement de les tecnologies que s'utilitzaran pot provocar que s'hagi d'invertir més temps aprenent-les, tot i que la immensa quantitat d'informació que hi ha a internet fa que l'impacte sigui reduït.

Per altra banda, el punt crític del projecte és l'obtenció de les dades de contaminació. Aquest punt és l'únic que depèn d'altres, ja que CommSensum no disposa de dades pròpies oficials. En cas d'haver-hi problemes s'hauran de cercar fonts alternatives de dades, tot i que el desenvolupament del crawler es veuria greument afectat i podria provocar una prolongació del projecte.

2.4.3 Recursos

En aquesta secció es detallen els recursos de *hardware* i *software* i els recursos humans que s'han utitlitzat durant la realització del projecte.

2.4.3.1 Recursos hardware

- Ordinador de sobretaula: Dell Optiplex 7040 Small Form factor
- Ordinador portàtil: MSI CX61
- Servidor dedicat de 4 cores i 32GB de memòria RAM

2.4.3.2 Recursos software

- Sistema operatiu: Debian 8.0 "Jessie", Ubuntu 16.04 i Windows 10 Education
- Editor de text: vim
- Control de versions: git / BitBucket
- Gestió del projecte: Trello
- Redacció de documents: Microsoft Office 2016 i Microsoft Project 2016
- Gestor de referències: Mendeley

2.4.3.3 Recursos humans

Podem diferenciar tres perfils:

- Cap de projecte
- Programador
- Administrador de sistemes

2.5 Pressupost

El cost del projecte ha acabat essent superior a l'estimat degut a la falta de dades que s'ha mencionat anteriorment. En aquesta secció s'hi mostren el pressupost inicial i el final dividits en els costs de recursos humans, *software* i *hardware*, i també els costos indirectes generats amb l'execució del projecte.

2.5.1 Estimació inicial del pressupost

En començar el projecte es va fer una estimació del pressupost que va quedar reflectida en les entregues inicials. A continuació s'hi inclou el pressupost total.

Concepte	Cost (€)
Cost de recursos humans	17.153,65
Cost de *hardware*	209,61
Cost de *software*	73,35
Costs indirectes	163,17
Contingència	5%
Subtotal	**18.479,77**
Imposts	21%
TOTAL	**22.360.52**

Taula 3. Estimació inicial

2.5.2 Pressupost final

2.5.2.1 Recursos humans

Tot i que el projecte el realitza una sola persona, haurà d'exercir els tres rols descrits anteriorment. La taula 4 mostra el pressupost dividit per tasca, mentre que la taula 5 mostra el pressupost dividit pels rols. Els preus s'han obtingut a partir dels estudis de MichaelPage[16] i PagePersonnel[17].

Fase	Temps dedicat (h)			Cost (€)
	Cap de projecte	Administrador de sistemes	Programador	
Planificació	75			3.012,00
Formació			80	2.034,40
Anàlisi i documentació del codi existent			35	890,05
Desenvolupament dels crawlers	40		130	4.912,30
Adquisició del servidor		40		1.269,60
Servidor intermediari de dades		70		2.221,80
Millores a la API de CommSensum			30	762,90
Documentació i memòria	100			4.016,00
TOTAL	**215**	**110**	**275**	**19.119,05**

Taula 4. Recursos humans

Rol	Preu per hora (€)	Hores	Total (€)
Cap de projecte	40,16	215	8.634,40
Administrador de sistemes	31,74	110	3.491,40
Programador	25,43	275	6.993,25
TOTAL		**600**	**19.119,05**

Taula 5. Recursos humans (2)

2.5.2.2 Hardware

Durant el projecte s'han utilitzat els recursos *hardware* que s'especifiquen a la taula següent. Es consideren un ordinador de sobretaula per treballar a la universitat, juntament amb el seu monitor; per treballar des de casa s'utilitzarà un ordinador portàtil; i un servidor dedicat de l'empresa OVH per allotjar-hi els serveis, concretament un model HOST-32L, amb un Intel Xeon D1520, amb 4 cores i 8 threads, 32GB de memòria RAM, 4TB d'emmagatzematge i una connexió a la xarxa de 1Gbps. El servidor s'ha adquirit per un període de sis mesos, que és el que es veu reflectit a la taula.

Es considera que la vida útil del *hardware* és de quatre anys, i que cada any té 250 dies hàbils amb jornades de 6 hores diàries. El cost d'amortització,també reflectit a la taula, es calcula com $\frac{preu}{vida\ útil * 250\ dies\ hàbils * 6 hores/dia}$.

Producte	Preu (€)	Vida útil	Cost d'amortització (€/h)	Amortització (€)
Dell Optiplex 7040 Small Form factor	1263,08	4 anys	0,2115	126,90
Monitor Dell P2414H	269,00	4 anys	0,0448	26,88
MSI CX61	749,00	4 anys	0,1248	74,88
Servidor OVH	359,94	N/A	N/A	359,94
TOTAL				**588,60**

Taula 6. Recursos hardware

2.5.2.3 Software

La majoria del *software* utilitzat és lliure i gratuït, de manera que no incrementa el cost del projecte. Es contemplen dues llicències, una de Microsoft Office i una de Microsoft Project, per elaborar la documentació.

Es considera que la vida útil del *software* és de tres anys, i el cost d'amortització es calcula de la mateixa manera que per al *hardware*.

Producte	Preu (€)	Vida útil	Cost d'amortització (€/h)	Amortització (€)
Debian 8.0	0,00	3 anys	0,00	0,00
Ubuntu 16.04	0,00	3 anys	0,00	0,00
Windows 10 Education	0,00	3 anys	0,00	0,00
vim	0,00	3 anys	0,00	0,00
Bitbucket	0,00	3 anys	0,00	0,00
Trello	0,00	3 anys	0,00	0,00
Microsoft Office 2016	123,14	3 anys	0,0274	16,44
Microsoft Project 2016	769,00	3 anys	0,1709	102,54
Mendeley	0,00	3 anys	0,00	0,00
TOTAL				**118,98**

Taula 7. Recursos software

2.5.2.4 Costs indirectes

El projecte es desenvoluparà a la universitat, de manera que no es comptabilitzen despeses d'internet però sí de desplaçament diari. El consum de corrent es calcula a partir del consum dels ordinadors i de 32 fluorescents de 20W que hi ha a la sala.

Producte	Preu	Unitats	Cost estimat (€)
Electricitat	0,12€/kWh	848W * 600h	61,06
Transport	28€/mes	7 mesos	196,00
Paper	5€	1 paquet	5,00
TOTAL			**262,06**

Taula 8. Costs indirectes

2.5.2.5 Cost total

El cost total del projecte es veu reflectit a la taula següent. No s'hi inclou un pressupost de contingència perquè aquest pressupost s'ha calculat una vegada finalitzat el projecte.

Concepte	Cost (€)
Cost de recursos humans	19.119,05
Cost de *hardware*	588,60
Cost de *software*	118,98
Costs indirectes	262,06
Subtotal	**20.088,69**
Imposts	21%
TOTAL	**24.307,31**

Taula 9. Cost total

2.5.3 Control de gestió

Una vegada acabat el projecte es pot calcular la desviació que hi ha hagut respecte el pressupost inicial. L'increment del cost és d'esperar donat que el projecte finalment ha tengut una durada superior a l'estimada inicialment, si bé el pressupost de contingència ha evitat que el sobrecost sigui excessiu. A més, al sobrecost per l'increment de temps s'hi ha d'afegir la inclusió del servidor dedicat en el pressupost de *hardware*, ja que inicialment no es contemplava la seva compra.

Concepte	Pressupost inicial	Cost final	Diferència
Recursos humans	17.153,65	19.119,05	1.965,40
Hardware	209,61	588,60	378,99
Software	73,35	118,98	45,63
Costs indirectes	163,17	262,06	98,89
Contingència	879,99	0,00	-879,99
Imposts	3.880,75	4.218,62	337,87
TOTAL	**22.360,52**	**24.307,31**	**1.946,79**

Taula 10. Sobrecosts

3. Sostenibilitat i compromís social

3.1 Sostenibilitat ambiental

Tenint en compte que aquest projecte es basa en el desenvolupament de *software*, la petjada ambiental que pot tenir es limita, principalment, al consum energètic degut al funcionament dels ordinadors i la il·luminació.

S'ha calculat que el consum elèctric és d'aproximadament 600kWh, que, utilitzant un factor de conversió de 1kWh ↔ 0.308kg de CO_2, equivalen a 184.8kg de CO_2. El desplaçament sempre es realitza amb transport públic, i la climatització del despatx està sempre encesa i només es pot modificar lleugerament, de manera que no tenen impacte en el projecte.

A més, el fet de reaprofitar l'aplicació d'airACT ha permès reduir l'impacte ambiental, evitant que es destinin recursos i temps a desenvolupar una aplicació que ja existeix.

Si es tornàs a començar el projecte tal vegada es podria reduir la petjada ambiental si s'aconseguissin les dades abans, ja que la durada del projecte seria menor. De totes maneres, si bé l'escala de temps es reduïria els càlculs de potència seguirien sent bastant semblants.

Per altra banda, la petjada ambiental durant la vida útil del projecte serà únicament la del *hardware* on s'executi, en aquest cas el servidor dedicat. Si bé el propietari no facilita

el consum real de la màquina, per les característiques s'estima que té un consum de 148W, que es tradueixen en 745,92kWh cada mes si està encès les 24 hores del dia.

Tot i aquest consum, l'impacte ambiental a nivell global és positiu, ja que una major exposició de la problemàtica de la contaminació atmosfèrica podria propiciar noves campanyes de pressió, que podrien acabar en una reducció de les emissions i un aire més net.

Respecte als possibles riscs, l'únic que es contempla és si el proveidor fes modificacions en el *hardware* que es traduïssin en un augment del consum. Aquest canvi quedaria fora de les nostres mans, i una alternativa seria buscar un altre servidor on allotjar-hi el projecte.

3.2 Sostenibilitat econòmica

Tal i com s'ha mostrat, el cost del projecte que es va estimar inicialment ha acabat essent insuficient, principalment per la falta de dades. Aquest problema era un risc que es coneixia i es tenia en compte, però no es va considerar que la falta de dades arribés a paralitzar el projecte durant tant de temps. Finalment hi ha hagut un sobrecost de quasi 2000 euros, tot i que si es té en compte el temps que s'ha allargat el projecte no és massa elevat. De totes maneres, si es deixa de banda la durada del projecte, el fet d'ampliar un projecte existent en lloc de començar-ho tot de nou ha evitat que el pressupost fos inclús superior, ja que hauria obligat a desenvolupar un *front-end* de bell nou.

Durant la vida útil d'aquest projecte s'haurà de mantenir un servidor on allotjar-lo. Si es segueix utilitzant el mateix servei el cost serà de 59,99€ més imposts cada mes. Si les necessitats del grup disminuïssin o si sorgís alguna oferta nova llavors es podria optar a un servidor més barat, reduint-ne el cost. A més, caldrà disposar d'alguna persona que s'encarregui de fer el manteniment del servidor i les màquines virtuals que s'hi allotgen per assegurar-se de que estan actualitzades i que no es queden sense espai al disc. Per la seva senzillesa, aquesta tasca només ocuparà unes hores i pot encarregar-se'n un membre del grup, de manera que no causaria una despesa addicional.

Un risc que cal tenir en compte és que les dades que actualment es reben des del servei de la EEA canviïn de format al llarg de la vida útil del projecte. Si això ocorregués s'hauria de disposar d'una persona que fos capaç de solucionar-ho i provocaria una despesa imprevista, tot i que el nombre d'hores hauria de ser reduït. En aquest cas es suposa un perfil de programador amb la mateixa retribució que s'ha usat anteriorment, 25,43€/h.

3.3 Sostenibilitat social

Aquest projecte no pretén ser una sol·lució innovadora, sinó resoldre un problema local i millorar l'oferta d'aquest tipus d'aplicacions a la regió. És per això que l'impacte que pot tenir sobre la societat és més limitat que una aplicació a nivell mundial, si bé la força d'Ecologistas en Acción, l'organització que va impulsar airACT, pot aconseguir que la repercussió sigui més gran. Amb l'extensió de les zones monitoritzades i, amb el temps, una xarxa de nodes CAPTOR en mans de la població es pot aconseguir una major

conscienciació en la qualitat de l'aire i la quantitat d'emissions nocives que hi ha avui en dia, fet que podria conduir a una reducció d'aquestes emissions.

A més, cal recordar que el gestor i beneficiari de l'altra cara del projecte és un grup de recerca de la universitat. Això pot significar que aquestes dades ajudaran en la recerca en el camp de l'IoT, un camp que té molt de futur i que podria significar un augment considerable del nivell de vida d'aquí a uns anys.

A nivell personal, aquest projecte ha estat una porta d'entrada al món la recerca universitària, un món que m'era totalment desconegut. També m'ha fet conèixer la problemàtica de l'ozó i els efectes nocius de la contaminació atmosfèrica. Així mateix, m'ha despertat una curiositat creixent en l'IoT i el tractament de dades, fins al punt de plantejar-me continuar la meva formació en aquest camí amb algun dels màsters MIRI de la UPC.

Finalment, no s'ha identificat cap risc social d'aquesta aplicació. L'existència d'altres aplicacions similars facilitaria que, en cas de desaparèixer, els usuaris poguessin canviar d'aplicació, amb l'única diferència de la localitat i l'abast de les dades. A més, en un món on l'*open data* està creixent i amb cada vegada més administracions públiques recolzant projectes d'aquest tipus, les fonts de dades seran de cada vegada més abundants, i facilitaran el desenvolupament d'aplicacions similars en un futur.

3.4 Matriu de sostenibilitat

Tenint en compte l'exposat en els punts anteriors, s'ha assignat una puntuació a cada casella de la matriu de sostenibilitat que es mostra a continuació. La primera columna representa el projecte prosat en producció (PPP), que és el que comprèn el TFG. Inclou la planificació, el desenvolupament i la implantació del projecte. La segona columna és la vida útil del projecte, des de la seva implantació fins a la seva retirada. La tercera columna representa els riscs del projecte des de la planificació fins a la retirada.

Havent assolit una puntuació de 66 sobre 90, podem dir que el projecte és sostenible, i que ho és en els àmbits en els que s'ha avaluat.

	PPP	**Vida útil**	**Riscs**
Ambiental	Consum del disseny	Petjada ecològica	Riscs ambientals
	8	17	-2
Econòmic	Factura	Pla de viabilitat	Riscs econòmics
	7	15	-5
Social	Impacte personal	Impacte social	Riscs socials
	9	18	-1
Rang sostenibilitat	24	50	-8
	66		

Taula 11. Matriu de sostenibilitat

4. Tecnologies implicades

A continuació es descriuen les tecnologies que s'han utilitzat al projecte. El fet de ser un grup de recerca de la universitat implica que hi passen molts d'estudiants i normalment en períodes relativament curts. S'ha optat sempre que ha estat possible per tecnologies que ja s'usen al grup de recerca, per així facilitar el treball dels professors i la incorporació d'aquests estudiants.

4.1 Python

Python és un llenguatge de programació d'alt nivell de propòsit general i un dels més usats actualment[X]. Les seves característiques principals són el tipatge dinàmic i la indentació forçada, i el seu disseny fa que el codi sigui fàcilment llegible. La popularitat de la que gaudeix ha donat lloc a una gran comunitat i una gran quantitat de llibreries. Aquest fet, sumat a la seva senzillesa, l'han convertit en un llenguatge molt popular per al tractament de grans quantitats de dades i com a llenguatge de *scripting*.

Actualment a la versió 3.6, els canvis en el salt de la versió 2.7 a la 3 varen provocar tal polèmica que actualment la versió 2.7 es segueix mantenint. Aquest projecte s'ha desenvolupat en Python 2.7, ja que és la versió en què estan escrits alguns dels programes del grup.

4.2 Hipervisor

Un hipervisor, també anomenat monitor de màquines virtuals, és una plataforma de *software* que permet aplicar tècniques de virtualització per executar diversos sistemes operatius sobre una mateixa màquina. La virtualització és un mecanisme que permet aïllar el *hardware* de la màquina del sistema operatiu, de manera que totes les crides al maquinari han de passar a través de l'hipervisor.

Existeixen dues classes d'hipervisor. Els hipervisors nadius o de tipus 1 són els que s'executen directament sobre el *hardware* de la màquina per controlar-lo i admnistrar les màquines virtuals. Els hipervisors hostatjats o de tipus 2 són els que s'executen a sobre d'un sistema operatiu convencional com un procés més, i creen una capa d'abstracció per a les màquines virtuals. La imatge següent mostra en un diagrama el funcionament d'un hipervisor.

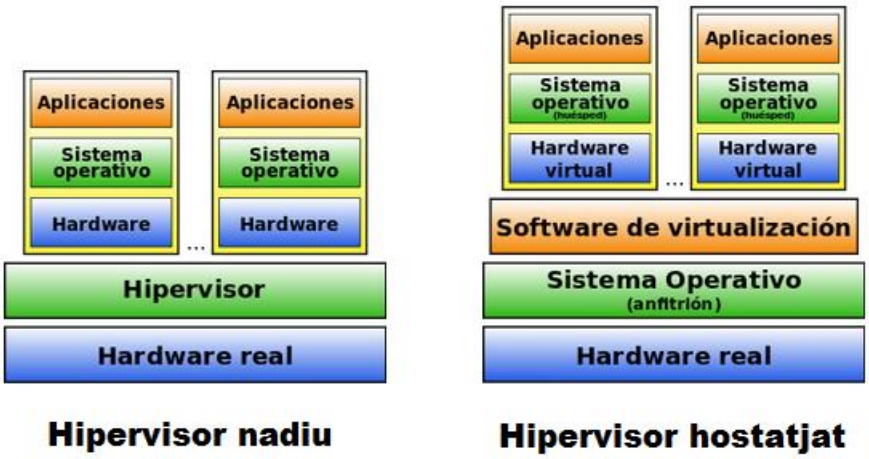


Figura 15. Diferències entre hipervisors

4.3 Citrix XenServer

XenServer és un hipervisor nadiu comercialitzat per Citrix basat en XenProject i és una de les opcions que ofereix OVH per instal·lar als servidors. La versió instal·lada és la 6.5, tot i que actualment ja ha sortit la versió 7.1. Citrix compta també amb XenCenter, un gestor de màquines virtuals gràfic per a Microsoft Windows, extensible amb *plugins*, que facilita l'administració de màquines virtuals. Per a plataformes GNU/Linux existeix OpenXenManager, un clon de codi lliure de XenCenter multiplataforma.

4.4 Ubuntu

Ubuntu és una distribució de GNU/Linux publicada per l'empresa Canonical Ltd. Està basada en Debian, i compta amb un repositori de *software* més actualitzat que el d'aquesta distribució. En aquest projecte s'ha optat per instal·lar la versió Ubuntu Server 16.04. Les versions de servidor es caracteritzen per ocupar menys espai que les d'escriptori, ja que no inclouen programari innecessari per a un servidor com pot ser una interfície gràfica. La versió 16.04, a més, és una versió amb suport a llarg termini (*Long-Term Support,* LTS en anglès), i per tant rebrà actualitzacions de seguretat fins el 2021.

4.5 SSH

SSH, de l'anglès *Secure Shell*, és un protocol que crea un canal de comunicació segur sobre xarxes insegures per connectar un client SSH amb un servidor SSH. Per a fer-ho utilitza criptografia de clau pública, tot i que permet la connexió amb contrasenya generant automàticament un parell de claus. S'ha acceptat com a estàndard de l'IETF, i se li ha assignat el port estàndard 22 de TCP. Tot i que va començar com a programari lliure, va anar evolucionant cap a llicències propietàries.

OpenSSH, un *fork* de l'última versió lliure del codi és avui en dia la implementació més estesa i ve instal·lat per defecte en la majoria de distribucions de GNU/Linux. És precisament aquesta versió la que s'utilitza per a connectar els nodes de CAPTOR al servidor.

4.6 PostgreSQL

PostgreSQL, o Postgres, és un sistema de gestió de bases de dades relacionals de codi lliure nascut el 1996. Les seves principals característiques són, en primer lloc, l'alta concurrència, és a dir, que permet que els usuaris llegeixin una taula mentre un altre usuari hi està escrivint sempre mantenint la consistència de la base de dades. En segon lloc, la gran varietat de tipus de dades de què disposa i, per últim, l'ús del llenguatge SQL estàndard (SQL:2008).

El *crawler* desarà les dades en una base de dades PostreSQL, concretament de la versió 9.5, que introdueix millores que faciliten l'execució del codi.

4.7 XML

L'XML, de l'anglès *eXtensible Markup Language*, és un llenguatge de *markup* d'estàndard obert desenvolupat per el *World Wide Web Consortium* (W3C) àmpliament utilitzat en l'intercanvi d'informació a internet. Es caracteritza per l'estructura en arbre i l'ús d'etiquetes, que permeten desar dades de forma organitzada i representar qualsevol estructura de dades de manera senzilla.

En el cas d'aquest projecte, per exemple, les dades de contaminació europees es reben i processen en format XML.

4.8 CSV

CSV, de l'anglès *Comm Separated Values*, és un format no estandarditzat per emmagatzemar dades en forma de taula en text pla. Les files estan separades per salts de línia, mentre que les columnes es separen per comes, tot i que també s'accepten altres caràcters per fer la funció de separadors, com el punt i coma.

És un format fàcilment llegible per programes d'edició de documents i fulls de càlcul, i el seu tractament és molt senzill. Tot i això, el fet de no estar estandarditzat obliga a definir els separadors des del principi per evitar que es mesclin diferents formats en un mateix document.

4.9 HTTP

De l'anglès *HyperText Transfer Protocol*, HTTP és l'estàndard més utilitzat en el *World Wide Web* per a la transferència de text i fitxers multimèdia. És un protocol de la capa d'aplicació que funciona amb un esquema de petició-resposta en un model client-servidor. Les peticions més utilitzades són GET i POST, tot i que n'hi ha d'altres com PUT, DELETE o HEAD.

HTTP és el protocol utilitzat en els serveis web REST, que permeten desenvolupar aplicacions distribuïdes on s'accedeix als recursos a través de peticions HTTP a adreces URI (*Uniform Resource Identifier*).

4.10 Node.js

Node.js és un motor d'execució de JavaScript de codi obert i multiplataforma que executa codi JavaScript al servidor, en lloc de fer-ho al costat del client, com s'ha fet

tradicionalment. Utilitza un sistema d'entrada/sortida asíncrona per reduir el temps d'espera de les peticions i incrementar l'escalabilitat del servei.

5. Desenvolupament

En aquest apartat s'expliquen en detall els elements que s'han desenvolupat durant el projecte. Com que el projecte s'executarà damunt el *hardware* adquirit, en primer lloc s'explicarà tot el procés de compra i creació de les màquines virtuals. Seguidament es detallarà el funcionament del *crawler* i de l'adquisició de les dades, per acabar amb la implementació del servidor intermediari de CSV. A la figura 16 s'hi mostra un diagrama de com estan interconnectades les parts del projecte entre elles i amb CommSensum i airACT.

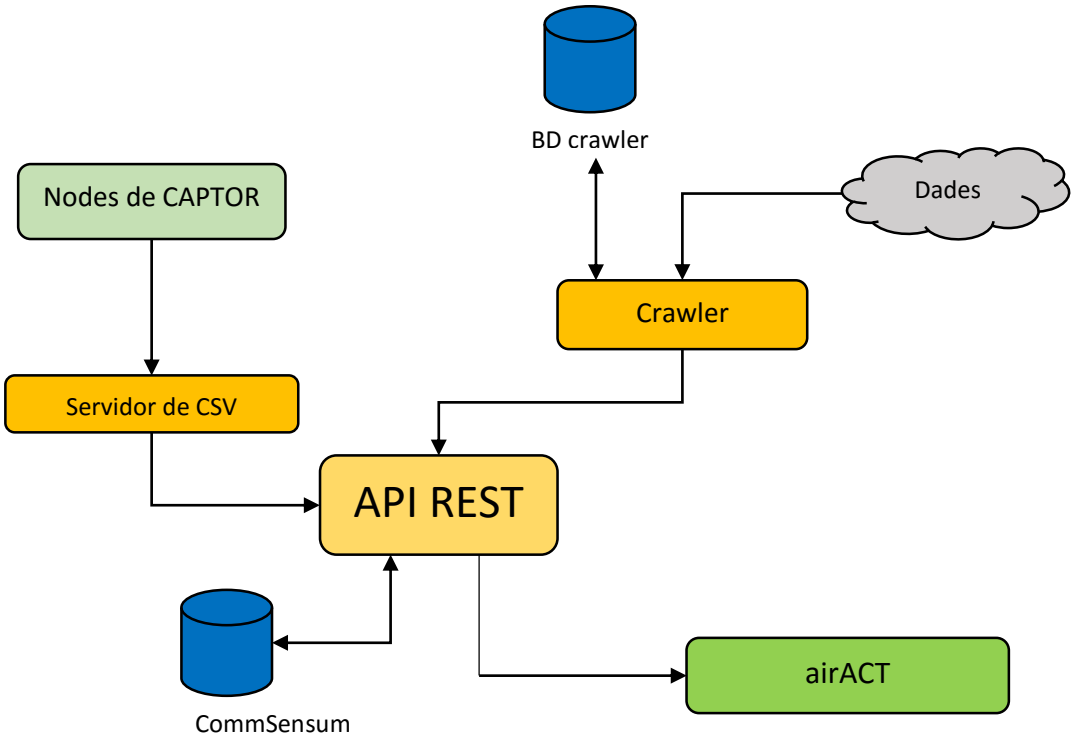


Figura 16. Funcionament del projecte

5.1 Adquisició d'un nou servidor

És important en un projecte informàtic tenir un maquinari adequat a les necessitats reals, que sigui capaç de suportar la càrrega de treball necessària. Quan aquest projecte va començar, CommSensum s'executava sobre un servidor del departament d'Arquitectura de Computadors amb un sol processador, 2GB de RAM i quasi 60GB d'espai d'emmagatzematge.

Aquestes característiques eren justes per executar l'API i la web de CommSensum, a més de recollir dades de Catalunya i Madrid. Una petició que implicàs fer més d'un SELECT a la base de dades causava una cua de peticions, i la resposta era, sovint, molt lenta.

5.1.1 Comparativa d'ofertes

Era palès que si s'havien de crear nous serveis aquest maquinari seria del tot insuficient, de manera que es feia necessari un *hardware* més potent. A la taula següent s'hi mostra una petita comparativa de preus de servidors de distints proveïdors. S'ha optat per mantenir la moneda original per evitar els canvis de preu produïts per la fluctuació dels mercats.

Empresa	Característiques	Preu
DigitalOcean	• 2 cores • 4GB RAM • 60GB SSD	40€/mes
GoDaddy Premium	• 4 cores • 32GB RAM • 2TB HDD	152,45€/mes (395,66 quan es renova)
RackSpace Starter	• 6 cores • 32GB RAM • 15TB HDD	449$/mes
OVH HOST-32L	• 4 cores/8 threads • 32GB RAM • 4TB HDD	59,99€/mes
Microsoft Azure A4m v2	• 4 cores • 32GB RAM • 40GB HDD	163,13€/mes (0,219€/hora)
Amazon EC2 t2.2xlarge	• 8 cores • 32GB RAM • Emmagatzematge contractat a part	0,376$/h
Amazon EC2 t2.medium	• 2 cores • 4GB RAM • Emmagatzematge contractat a part	0,047$/h

Taula 12. Comparativa d'ofertes

Com es pot veure, els preus són molt variables. Les quatre primeres empreses ofereixen servidors dedicats, és a dir, servidors on es disposa de tot el *hardware* per al seu ús. Les ofertes de Microsoft i Amazon, per altra banda, són de servidors compartits, on l'usuari disposa d'una màquina virtual i comparteix els recursos de hardware amb altres usuaris.

El cas d'Amazon EC2 mereix especial atenció. El preu d'aquestes màquines es calcula segons les hores de CPU que es consumeixen, i el preu per hora varia segons la instància contractada, tal i com es pot apreciar a la taula. Per això contractar una instància t2.x2large per instal·lar-hi un hipervisor no tendria gaire sentit, ja que estaríem introduint un consum de CPU innecessari amb un preu superior: amb un ús de només 6 hores diàries la despesa es dispara a 67,68$ cada mes.

L'alternativa seria contractar tantes instàncies t2.medium com fossin necessàries. Aquestes instàncies facturarien un màxim de 33,84$ al mes, amb un consum de CPU de vint-i-quatre hores al dia, un cas al que no s'hi arribaria mai. Amb el mateix consum de 6 hores diàries que l'exemple anterior la factura seria de 8,46€, quantitat a la que se li ha d'afegir el preu de l'emmagatzematge, 0,045€/GB-mes.

Tot i això, el consum i, en conseqüència, la factura són variables, i s'ha d'estudiar com escala el servei amb la inclusió dels nous projectes per a poder fer una previsió del consum.

Per això el major problema a l'hora de contractar un servidor és la burocràcia. Tots els pagaments els ha de fer la universitat, i és molt més senzill de justificar i acceptar un preu fix de 59,99€ cada mes, amb la possibilitat de contractar sis mesos o un any en un sol pagament, que una factura variable cada mes.

Per aquesta raó al final es va decidir contractar el servidor dedicat d'OVH. Aquest servidor, a més de les característiques reflectides a la taula 12, proporciona un ample de banda de 250Mbps amb un augment automàtic de fins a 1Gbps per a necessitats puntuals. A més, en cas de contractar-hi un altre servidor disposaríem d'una connexió entre ells d'1Gbps. També ofereix fins a 256 adreces IP sense cost addicional, un espai per a fer-hi *backups* de 500GB i suport en cas de problemes. Per últim, garanteix una disponibilitat del 99,982%, és a dir, que garanteix un màxim d'1,6 hores d'interrupció del servei cada any. Aquesta característica és especialment atractiva per a un projecte que rep dades de nodes IoT constantment.

5.1.2 Instal·lació de l'hipervisor i creació de màquines virtuals

Una vegada es té accés a la consola d'administració, instal·lar un sistema operatiu és molt senzill. Només cal seleccionar un dels *templates* que ofereix OVH i fer clic a instal·lar. OVH ofereix molts de *templates* diferents. Entre d'altres, permet triar entre diverses distribucions de GNU/Linux o Unix, com Debian, Ubuntu, CentOS, Gentoo, OpenSUSE, Fedora o FreeBSD; diverses versions de Windows Server, en concret 2008, 2012 i 2016; i diversos hipervisors de tipus 1, com VMware ESXi, Microsoft Hyper-V o Citrix Xen Server.

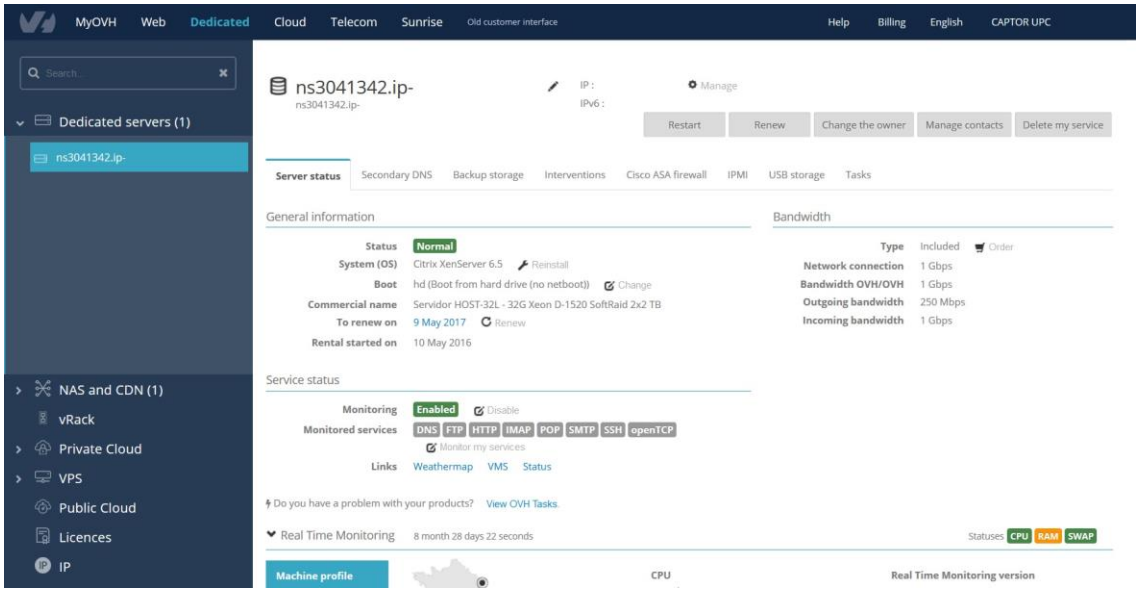


Figura 17. Panell de control d'OVH

Per a aquest projecte s'ha decidit utilitzar Citrix Xen Server 6.5, una versió que s'ofereix de forma gratuïta i és senzilla d'utilitzar. Una vegada instal·lat, l'administració de

màquines virtuals es pot fer per línia de comandes o, més fàcilment, utilitzant el gestor que proporciona Citrix: XenCenter.

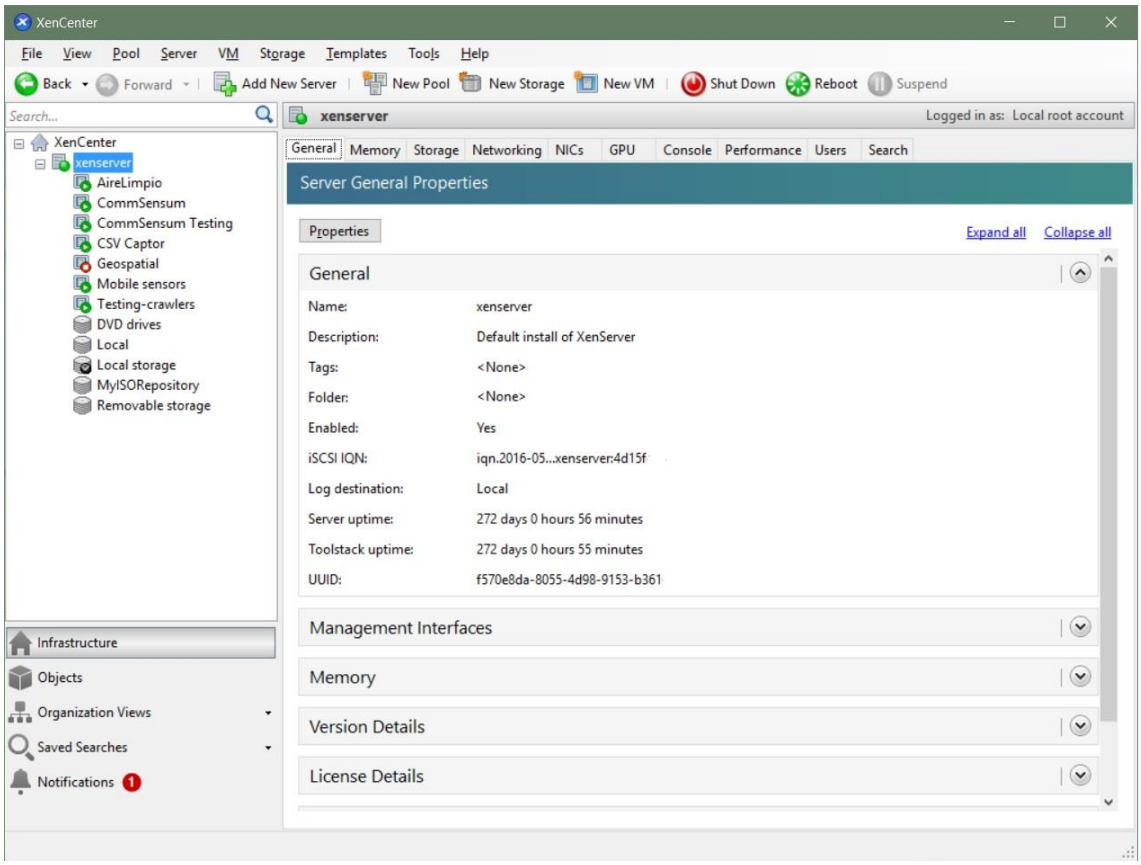


Figura 18. XenCenter

Per a crear una màquina virtual nova, només s'ha de clicar a *New VM*, que ens obrirà una finestra nova. Després de triar un *template* per a la màquina i d'assignar-li un nom, podem triar si insta·lar el sistema operatiu des d'un disc local o a través de la xarxa. En aquest cas, com que la versió d'Ubuntu del template és una versió d'escriptori ens hem decantat per instal·lar-la amb una imatge de disc local.

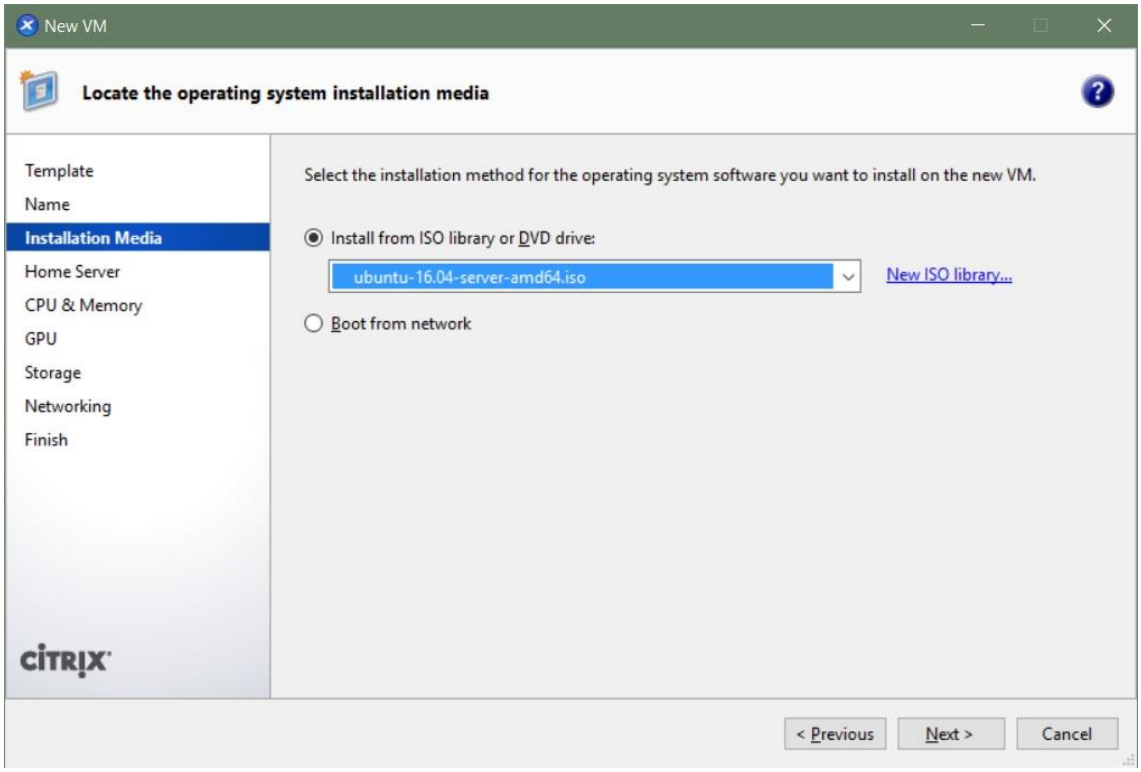


Figura 19. Creació d'una màquina virtual

Una vegada assignats els recursos que considerem necessaris s'ha de configurar la xarxa, el punt menys intuïtiu de tota la instal·lació. Per sort, OVH compta amb tutorials que guien pas a pas durant la configuració de la xarxa. És necessari haver assignat una IP de les disponibles a una adreça MAC virtual a la consola d'administració d'OVH. Aquesta MAC virtual serà la que assignarem a l'interfície de xarxa de la màquina virtual.

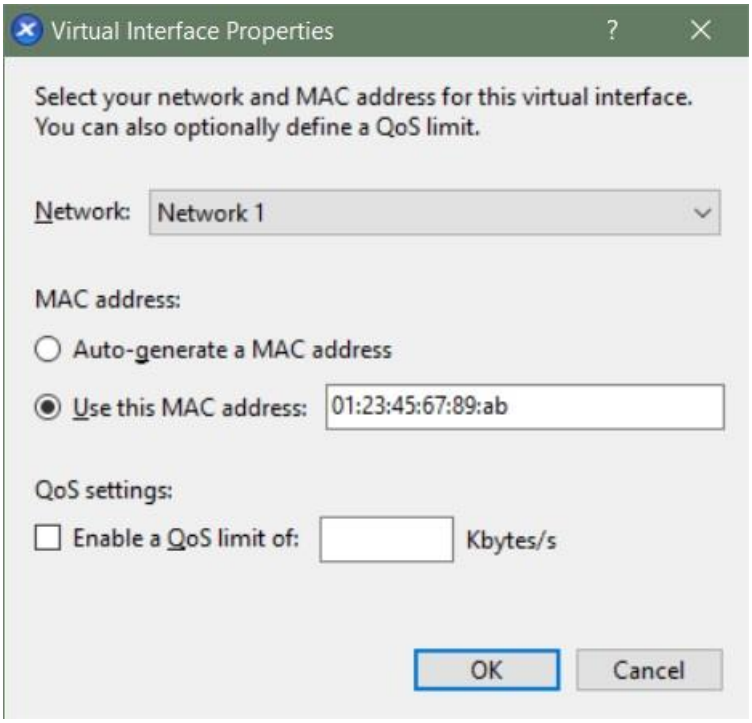


Figura 20. MAC virtual del la màquina

Per últim s'ha d'iniciar la màquina virtual i modificar la configuració de xarxa, que en una distribució basada en debian es troba a `/etc/network/interfaces`. Per a fer-ho es necessita disposar de la IP de la màquina i de la IP del *gateway*, que s'obté canviant l'últim octet de l'adreça del servidor dedicat per 254. Així, si l'adreça del servidor dedicat fos 123.56.189.101 i l'adreça que volem assignar a la màquina virtual fos 147.158.69.73:

```
iface eth0 inet static
        address 147.158.69.73
        netmask 255.255.255.255
        network 147.158.69.73
        broadcast 147.158.69.73
        post-up route add 123.56.189.254 dev eth0
        post-up route add default gw 123.56.189.254
        pre-down route del 123.56.189.254 dev eth0
        pre-down route del default gw 123.56.189.254
        dns-nameservers 213.186.33.99
```

5.2 Desenvolupament del *crawler*

El desenvolupament del crawler té dues part molt diferenciades: l'adquisició de les dades i la implementació del propi *crawler*. Per a la implementació s'ha decidit utilitzar Python, ja que és un llenguatge de programació senzill, fàcil d'aprendre i que serveix per a tant per desenvolupar aplicacions complexes com per fer *scripts*. Disposa d'un gran nombre de llibreries que faciliten l'escriptura del codi i, si bé el fet de ser un llenguatge interpretat implica que la majoria de vegades serà més lent que un llenguatge compilat, en aquest projecte es valora la simplicitat del codi per sobre de la velocitat d'execució. Per sobre de tot, però, és un llenguatge amb el que els professors del grup, que són els destinataris del projecte, ja hi estan familiaritzats.

És necessari que el *crawler* s'executi de forma periòdica per tenir les dades actualitzades. La solució més senzilla i adequada és crear una tasca al crontab per executar-lo cada hora. Qualsevol missatge de sortida s'emmagatzema com a mail a la carpeta `/var/mail`.

5.2.1 Obtenció de les dades

Tal i com ja s'ha explicat, el desenvolupament del *crawler* va patir el problema de manca de dades, que va provocar un retard al projecte.

5.2.1.1 Origen de les dades

Inicialment es varen voler utilitzar les dades que publiquen les distintes comunitats autònomes. Les dades de Catalunya i Madrid s'obtenien a través de l'*Open Data* d'aquestes comunitats, però la majoria de comunitats només publiquen les dades sobre el mapa de la regió, i sovint en una escala visual i informativa, però on hi manquen les dades numèriques. Es va provar de posar-se en contacte via correu electrònic amb els departaments de cada administració que s'encarreguen de la recollida, tractament i publicació de les dades, tot i que la resposta va ser escassa. De les comunitats que varen respondre, la immensa majoria no va oferir cap solució alternativa a la web que oferien, de manera que es va optar per cercar una altre origen de dades.

La segona opció, demanar les dades al Ministeri de medi ambient, tampoc va tenir èxit. ja que, després de dues setmanes a l'espera d'una resposta, se'ns va denegar l'ús de les dades que publicaven a la pàgina web per al projecte.

No va ser fins a l'estiu que es va trobar, a través d'un *partner* del projecte, el servei web de la *European Environment Agengy* (EEA). Aquestes dades es poden utilitzar sense problema, si bé s'ha de demanar un *token* d'accés que va trigar un temps a arribar.

Finalment es descarreguen les dades de pol·lució d'Espanya i Àustria. Actualment Itàlia, el tercer país on es despleguen nodes de CAPTOR, no disposa de dades horàries en temps real, de manera que es va haver d'optar per deixar-lo de banda.

5.2.1.2 Estructura de les dades

Les dades del servei de la EEA es descarreguen en format XML. Cada valor horari és un *record* a dins un array de *records*, i inclou informació diversa. En primer lloc, informació de la xarxa, que inclou el nom de la xarxa, el codi ISO de l'estat al que pertany i el fus horari on està situat. En segon lloc hi ha la informació de l'estació on s'ha recollit la mesura, amb un identificador de l'estació, així com un codi i el nom, i l'altitud a la que es troba. Seguidament es mostra la informació del sensor, anomenat *sampling point*, del que se'n dóna un identificador i les coordenades geogràfiques exactes, que ens permetran representar les dades sobre el mapa. A continuació es defineix el contaminant del que se n'ha pres la mesura. La EEA recull actualment informació de benzè, monòxid de carboni (CO), diòxid de nitrogen (NO2), òxids de nitrogen (NOX), ozó (O3), diòxid de sofre (SO2) i partícules en suspensió de diàmetres <10µm (PM10) i <2.5µm (PM2.5). Tot i això, no tots els països participants recullen mostres de tots els contaminants. Per últim, s'inclou la informació de la mesura: l'hora de la mesura, l'hora d'inserció a la base de dades, l'hora d'actualització —si ha calgut—, i el propi valor amb les unitats corresponents. A continuació es mostra un exemple corresponent a una mesura presa a les Illes Balears.

```
<?xml version="1.0" encoding="UTF-8"?>
<records>
<record>
<network_namespace>ES.BDCA.AQD</network_namespace>
<network_localid>NET_ES204A</network_localid>
<network_name>CCAA Islas Baleares</network_name>
<network_countrycode>ES</network_countrycode>
<network_timezone>http://dd.eionet.europa.eu/vocabulary/aq/timezone/UT
C</network_timezone>

<station_namespace>ES.BDCA.AQD</station_namespace>
<station_localid>STA_ES1604A</station_localid>
<station_code>ES1604A</station_code>
<station_name>BELLVER</station_name>
<station_altitude uom="m">117</station_altitude>

<samplingpoint_namespace>ES.BDCA.AQD</samplingpoint_namespace>
<samplingpoint_localid>SP_07040003_14_6</samplingpoint_localid>
<samplingpoint_point x="2.6205499999999997" y="39.563319999472519"
coordsys="EPSG:4979"/>

<pollutant>03</pollutant>

<value_datetime_begin>2017-04-11 04:00:00</value_datetime_begin>
<value_datetime_end>2017-04-11 05:00:00</value_datetime_end>
<value_datetime_inserted>2017-04-11 07:20:13</value_datetime_inserted>
<value_datetime_updated/>
<value_numeric unit="ug/m3">51</value_numeric>
<value_validity>1</value_validity>
<value_verification>3</value_verification>
</record>
...
</records>
```

5.2.2 Desenvolupament del *crawler*

El *crawler* està format per dos fitxers:

- **europa.py**: Aquest fitxer és el codi principal del crawler. Té diverses funcions, depenent de les opcions amb les que s'executi:
  - **-h**, **--help**: Mostra l'ajuda del programa
  - **--sql**: Genera un fitxer SQL que crearà l'esquema de taules per executar el programa
  - **-c codi**: Permet modificar els països dels que es descarregaran les dades introduint el codi ISO-3166, per defecte Espanya (ES) i Àustria (AT)
  - **-p contaminant**: Permet modificar els contaminants dels que es descarregaran dades introduint els codis de contaminant acceptats per la EEA.
  - **-u**: Descarrega les dades actualitzades durant l'última hora en lloc de les inserides.
  - **--send-ovh**: Envia les dades en cua al servidor.
  - **--send**: Versió anterior d'enviament que envia les dades en cua al servidor d'una manera més ineficient.

- **crawler_functions.py**: Aquest fitxer conté les funcions que es criden des d'europa.py i altres funcions internes.

5.2.2.1 Descàrrega de dades

La descàrrega de dades es fa a través del servei web de la EEA mitjançant una petició HTTP POST amb els següents paràmetres:

Nom	Descripció
UserToken	*Token* únic d'usuari necessari per accedir al projecte. Si el *token* no és vàlid no es retornen dades.
FromDate	Data d'inici de les dades, en format "yyyy-MM-dd hh:mm". Usat amb el paràmetre ToDate, el seu ús és opcional.
ToDate	Data de final de les dades, en format "yyyy-MM-dd hh:mm".
Countrycode	Codi ISO-3166 dels països dels que es volen descarregar dades, separats per una coma.
InsertedSinceDate	Filtra les dades segons l'hora a que s'han inserit a la base de dades de la EEA. També usa el format "yyyy-MM-dd hh:mm".
UpdatedSinceDate	Filtra les dades segons l'hora a que s'han actualitzat a la base de dades de la EEA. També usa el format "yyyy-MM-dd hh:mm".
Pollutant	Filtra els resultats segons el contaminant. Els noms dels contaminants han de ser els codis acceptats per la EEA[X].
NameSpace	Per filtrar per *namespace*.
Format	Format en el que s'enviaran les dades. Permet triar entre XML i tres formats diferents de CSV.
Qoute_output_values	Defineix si els valors s'envien entre cometes o no.
RunAsync	Permet decidir si utilitzar el nou servei de descàrrega de dades asíncron o si utilitzar el servei síncron tradicional.

Taula 13. Paràmetres de la EEA

El *crawler* s'executa dues vegades cada hora per recollir dades. La primera es descarrega les dades que s'han inserit durant les dues hores anteriors. La segona fa el mateix, però es descarrega les dades actualitzades. Aquesta diferència de dues hores no és arbitrària, és necessària per compensar el retard de les dades, que pot ser superior a una hora.

Una vegada s'han obtingut les dades, de cada *record* se n'extreuen les dades relatives a la comunitat autònoma, al node, al sensor i a la mesura i s'insereixen a la base de dades. S'ha optat per utilitzar una base de dades PostreSQL a petició dels professors del grup, més acostumats a utilitzar-la. Aquesta base de dades segueix el mateix esquema que la de CommSensum per simplificar la integració dins el grup, tot i que només s'hi han inclòs les taules necessàries per al funcionament del *crawler*.

Per enviar les dades a CommSensum s'ha decidit utilitzar un fitxer cua en format CSV, on s'escriurà cada mesura amb les dades necessàries per al seu enviament de la forma "DATA; nom sensor; comunitat; data; contaminant; valor; unitat de mesura"

A més, per a cada mesura es comprova si el node i el sensor ja existeixen a la base de dades. Si no és així, abans de la línia anterior s'escriuen dues línies amb el format "PROJECT; comunitat; comunitat; llicència; estat; usuari" i "NODESTREAM; comunitat; estació; descripció; id; latitud; longitud; altitud; nom sensor; descripció sensor; país".

5.2.2.2 Enviament de dades

Cada hora el *crawler* envia totes les dades que té en cua a CommSensum. Per cada línia del fitxer CSV es crida a una funció de l'API segons si comença per PROJECT, NODESTREAM o DATA. Per a connectar-se a l'API de CommSensum és necessari incloure l'usuari i el *token* únic a la capçalera de la connexió, mentre que el contingut del cos varia segons la funció. Totes les dades necesàries es llegeixen del fitxer cua.

En funció del resultat de l'operació, CommSensum retorna un codi d'estat HTTP. En cas de que aquest codi fos d'error, al final de l'execució es tornarà a escriure la fila al fitxer cua, pendent d'enviar-se al cap d'una hora. L'única excepció és l'error per entrada duplicada, que s'ignora.

La primera versió que es va desenvolupar utilitzava la versió tradicional de l'API, on cada dada s'envia a través d'una connexió HTTP. Tenint en compte que algunes hores hi podia haver més de 40.000 entrades pendents d'enviar, això provocava un col·lapse del servidor, que es bloquejava i refusava connexions. Introduir un retard en les peticions era inviable, ja que inclús amb un retard entre peticions de només 0,1 segons el programa tardaria més d'una hora en completar l'enviament.

Per a evitar aquesta situació es va actualitzar l'API de CommSensum per incloure-hi una funció que havia desenvolupat un company però que no havia estat provada en producció. Aquesta funció permet inserir diverses dades de cop en una sola petició. Això, juntament amb la naturalesa asíncrona de node.js, permet que l'enviament des de la part del client sigui molt ràpida, tot i que introdueix la possibilitat —remota, això sí— de que el servidor falli una vegada acceptades les dades, provocant-ne la pèrdua.

Per utilitzar aquesta funció totes les dades s'insereixen en un array que s'envia quan arriba a 100 elements. Aquest número és una mida arbitrària per sota del límit de 200 que té la funció, però és un bon equilibri entre la velocitat d'execució i el risc de pèrdua de dades.

5.2.2.2.1 Millora de l'API

CommSensum compta amb diversos mètodes per inserir dades, així com una funció que permet crear un node i un sensor a la base de dades, però no existia cap funció per crear un nou projecte.

Per tal de que aquest projecte pugui funcionar sense requerir una llista exhaustiva de tots el projectes que ha d'actualitzar, era necessari que tingués la capacitat de donar d'alta nous projectes. Tot i ser només una funció, el total desconeixement de node.js quan es va descobrir aquesta necessitat ha provocat que s'hi hagi dedicat més temps.

Aquesta funció s'utilitzarà per crear un nou projecte a la base de dades de CommSensum amb el nom de la regió —a Espanya, per exemple, comunitats autònomes—, i l'assignarà a l'usuari corresponent, en aquest cas l'usuari del *crawler*.

5.3 Construcció del servidor intermediari de dades

L'interès principal d'obtenir les dades de les estacions de referència és poder-les comparar amb les dades que obtenen els nodes de CAPTOR. Aquests nodes originalment enviaven les dades directament a CommSensum, però era necessari tenir-les desades en un lloc on fossin de més fàcil accés.

És per aquesta raó que es va decidir muntar un servidor intermediari. Els nodes de CAPTOR es connectarien al servidor i hi desarien les dades. Més tard, seria el propi servidor el que s'encarregaria d'enviar les dades a CommSensum. Això permet als nodes de CAPTOR reduir el consum de dades 3G, ja que actualment encara conviuen la versió de CommSensum que hi ha a la màquina del departament amb la versió que hi ha al servidor adquirit. D'aquesta manera només han d'enviar les dades una vegada, fet molt positiu en una connexió de dades limitada.

El format que s'ha triat per desar les dades al servidor és CSV. Aquesta decisió s'ha pres perquè els investigadors del CSIC que col·laboren amb el projecte CAPTOR necessiten accedir a les dades que es generen. Aquests investigadors són personal sense formació informàtica, i utilitzen fulls de càlcul per tractar les dades. El format CSV és un format llegible pels fulls de càlcul, al contrari que formats com XML o JSON. Per altra banda, els professors i alumnes del grup que utilitzen les dades usen Python per tractar-les, i és trivial llegir i escriure dades en un fitxer CSV.

5.3.1 Aïllament dels nodes

Com que els nodes CAPTOR es reparteixen a voluntaris de diversos països, era necessari introduir una capa de seguretat per evitar que una persona malintencionada pogués accedir al servidor a través del CAPTOR i modificar les dades d'altres nodes. En primera instància es va pensar en implementar una API, però aquesta opció es va descartar immediatament per la quantitat de treball que significaria. A més, construir una nova API implicaria haver de mantenir un nou projecte, amb un llenguatge o *framework* que quedaria desactualitzat d'aquí a pocs anys, i que obligaria a dedicar algú al seu manteniment.

La solució que es va triar és una solució que aprofita la gestió d'usuaris i permisos de GNU/Linux i les capacitats del servidor de SSH per aïllar els usuaris en el que es coneix com a *SSH Jail*.

La creació d'usuaris és senzilla. Els nodes de CAPTOR tenen noms amb números consecutius, de manera que es pot crear un *script* que els creï. Els nodes formaran part del grup "captor" i del grup "jail".

```
#!/bin/bash

addgroup captor
addgroup jail

for i in {00..60}; do
        user="captor170$i"
        echo $user
        adduser --quiet --home /home/$user --no-create-home --ingroup
captor --disabled-password --gecos "$user" $user
        adduser $user jail
done
```

Seguidament s'ha de crear l'espai on es connectaran. Un *SSH jail* fa ús de chroot per evitar que els usuaris surtin d'un espai controlat. Aquest espai serà `/home/jail/`. Aquí dins s'hi crearà un espai que simularà l'arrel del sistema, si bé només disposarà dels programes que necessitin els nodes. Per tant, és necessari crear els directoris `/var/root/{dev,etc,lib,usr,bin}`, així com `/var/usr/bin`. Aquests directoris, com l'arrel del sistema, han de ser propietat de root:root.

A continuació s'han de copiar els programes que es vulguin utilitzar a dins el *jail*. En aquest cas només hi inclourem `bash`, `cat`, `cp`, `echo` i `ls`. Aquests programes, però, no funcionaran sense les llibreries corresponents. Per trobar-les i copiar-les a dins el *jail* existeix un *script* anomenat l2chroot, inclòs a l'annex III. Només modificant la variable "BASE" s'encarrega de trobar i copiar totes les llibreries necessàries.

Per últim, cal modificar el fitxer de configuració del servidor SSH per a que enviï tots els usuaris del grup "jail" a la zona designada com a tal i els en restringeixi la sortida. Per fer-ho només s'han d'afegir al fitxer `/etc/ssh/sshd_config` les següents línies:

```
Match Group jail
        ChrootDirectory /home/jail
        X11Forwarding no
        AllowTcpForwarding no
        AuthorizedKeysFile /home/jail/home/%u/.ssh/authorized_keys
        PasswordAuthentication no
```

Aquesta configuració, a més, obliga que totes les connexions s'autentiquin amb parells de clau pública i privada. Això evitarà que un atac de força bruta pugui arribar a descobrir la contrasenya d'un usuari, afegint així una capa més de seguretat.

Per últim, cal assignar els permisos pertinents a cada directori. S'ha optat per establir cada usuari com a propietari del seu home, però el grup és "captor" per a tots. Els permisos són 740 per a cada directori, és a dir, que un usuari pot llegir i escriure en el seu directori i només pot veure quins fitxers hi ha en els directoris dels altres usuaris.

5.3.2 Enviament de les dades

L'enviament de les dades és gairebé idèntic al del *crawler*. Un *script* en Python llegeix els fitxers CSV que han escrit els nodes de CAPTOR i els envia a CommSensum.

El fitxer CSV varia una mica respecte al del crawler. En aquest cas s'utilitzen tabuladors com a delimitador, i les files contenen la data de la mesura, les mesures dels sensors del

node (els cinc sensors d'ozó, el de temperatura i el d'humitat), el càlcul de la mesura convertida a unitats convencionals i la data de calibració.

A més, els mou al directori de l'usuari csic, de manera que totes les dades estaran en un mateix directori, i per a un usuari no experimentat serà senzill accedir-hi amb programes com PuTTY o WinSCP.

6. Possibles millores

6.1 Inclusió de nous orígens de dades

La principal millora d'aquest projecte seria la inclusió de nous orígens de dades. Malauradament, la EEA encara no compta amb les dades d'Itàlia.Tot i això, hi ha *partners* italians del projecte que s'estan esforçant en aconseguir-les de les administracions locals. A més de les dades italianes, però, afegir nous països a l'aplicació d'airACT seria sens dubte positiu.

Si no s'ha contemplat en aquest projecte ha estat a causa de que es desconeix com escalarà l'aplicació i quina quantitat de dades es generarà. Una vegada resolts aquests dubtes el *crawler* està preparat per recollir noves dades i inserir-les a CommSensum.

De la mateixa manera, seria una gran millora aconseguir dades de més contaminants, si bé això depèn directament de les entitats que s'encarreguen de la recollida de dades.

6.2 Canvi de model de base de dades

Lligat amb el punt anterior, un creixement massiu de la quantitat d'informació podria causar una reducció del rendiment de la base de dades actual. Una possibilitat per solucionar-ho seria utilitzar una base de dades no relacional, com mongoDB o Cassandra, molt esteses avui en dia.

MongoDB en concret és una base de dades d'alt rendiment, escalable i sense esquema, és a dir, que permet inserir qualsevol tipus de dada. Aquesta característica podria ser un gran avantatge a l'hora d'incloure noves fonts d'informació que treballessin amb formats molt diferents de dades.

6.3 Monitorització dels nodes de CAPTOR

Una altra possible millora d'aquest projecte està relacionada amb el servidor de CSV. Aprofitant que els nodes s'hi connecten cada mitja hora per escriure-hi les dades, es podria escriure un *script* que monitoritzés quins nodes no s'han connectat, fet que podria indicar errors de connexió o problemes més greus que implicarien que el node no estàs recollint dades.

Actualment només es vigila si un node ha enviat alguna dada durant el dia per determinar que està operatiu. Ara bé, amb accés al servidor es pot fer un seguiment més exhaustiu dels nodes, i saber si estan operatius o detectar que les dades surten dels rangs normals de funcionament permet maximitzar-ne el temps de funcionament i de recollida de dades.

7. Conclusions

Aquest projecte va començar amb la intenció de recollir les dades de contaminació de diverses zones i, si bé l'objectiu principal s'ha mantingut, els objectius secundaris han variat al llarg del projecte. La duració d'aquest treball s'ha allargat molt més del que s'havia concebut en un principi. Aquest problema s'havia plantejat com a possible risc, però en cap moment es va imaginat que tindria la magnitud que va acabat tenint. Crec que va fer falta una millor valoració del risc, i que un coneixement més adequat dels organismes que recullen dades haurien pogut pal·liar el problema.

A nivell tècnic, crec que és important tenir en compte que si bé la tecnologia, i sobre tot la informàtica, evoluciona cada dia, no sempre el més nou és millor. Hi ha tecnologies que desapareixen al cap de poc d'haver nascut i suposa un problema per a tots els que l'han utilitzat. No tot necessita una solució amb una API d'última generació, si bé avui en dia la programació web pareix que és el camí que ha triat. És per això que s'ha optat per una solució "*low-tech*" per al servidor, una solució robusta en quant al pas del temps, que utilitza tecnologies estàndard conegudes per tothom.

Crec també que aquest projecte és un projecte amb futur. La contaminació atmosfèrica és un tema que, per sort, preocupa de cada dia a més gent, i una aplicació que mostri l'estat de l'aire que respiram té possibilitats de seguir amb vida. A més, el món de l'*Internet of Things* és un món que creix de cada vegada més, i d'aquí a uns anys podríem aconseguir que els particulars participin en el projecte CAPTOR. Amb tot, és difícil saber exactament quin impacte real tendrà aquest projecte sobre la societat. El millor que es pot fer és esperar que les previsions fetes a l'apartat de sostenibilitat es compleixin.

A nivell personal, valoro molt positivament l'experiència guanyada durant aquest projecte. Per una part, els coneixements que he adquirit són molts. He après llenguatges de programació, com Python i JavaScript, i he aprofundit el meu coneixement de Linux. Però, sobre tot, he après a treballar en equip molt més que en qualsevol assignatura del grau. La feina feta depèn d'altres, i altres depenen d'aquesta feina. Aquest fet, sumat a treballar en un despatx amb altres estudiants que realitzen projectes similars i interconnectats ha estat una de les experiències més positives del projecte.

Això m'ha fet veure també que un projecte no pot existir sol. Sempre es necessita la feina anterior d'una altra persona per treballar, i mai saps quan aquesta feina serà la teva. És per això que convé deixar-ho tot a punt per a que un altre pugui sentir-s'hi benvingut, com a mínim tant com t'agradaria trobar-ho a tu.

Annex I. Llicències de dades

Les dades que es descarrega el *crawler* de la EEA són lliures d'ús, tal i com recull la seva política de dades[X]. A continuació se n'inclou un fragment rellevant.

ARTICLE 4: ACCESS TO AND REDISTRIBUTION OF DATA

Access to data covers both technical access and the policies that govern access.

Products created by EEA are considered a public good and where possible, they will be made fully, freely and openly available for others to use.

All data held by EEA shall be made available with minimum time delay and at no cost except where

- restrictions apply resulting from binding rules, including international treaties, European Union law and national legislations including the protection of personal data, statistical confidentiality, the protection of intellectual property rights as well as the protection of national security (i.e. State security), defense, or public security,
- data made available by EEA is accompanied by a data license. Data originally made available to EEA by a third-party may have their own data access agreements and license conditions agreed upon with EEA, which restricts how or when EEA can make data available to others,
- the data access request exceeds EEA's handling capacities.

Tot i això, convé sempre reconèixer l'origen de les dades, i així s'ha fet a airACT:

https://airact.org/about/

Annex II. Desplegament del *crawler*

Per desplegar el *crawler* primer s'ha d'actualitzar el software de la màquina. Per fer-ho cal executar

```
sudo apt-get update
sudo apt-get upgrade
```

El projecte assumeix que es treballa amb una versió de PostgreSQL igual o superior a la 9.5, ja que d'altra manera les insercions a la base de dades resultaran en error per l'ús de característiques que es varen introduir en aquesta versió.

Seguidament s'ha d'obtenir el codi:

```
git clone https://murq@bitbucket.org/murq/crawlers.git
```

Per instal·lar les dependències del projecte cal executar

```
cd crawlers
pip install –r requirements.txt
```

I per últim, executar el crawler:

```
python europa.py
```

Annex III. l2chroot

A continuació s'inclou el *script* l2chroot, usat en la configuració del *SSH jail*.

```
#!/bin/bash
# Use this script to copy shared (libs) files to Apache/Lighttpd
chrooted
# jail server.
# ------------------------------------------------------------------------
--------
# Written by nixCraft <http://www.cyberciti.biz/tips/>
# (c) 2006 nixCraft under GNU GPL v2.0+
# + Added ld-linux support
# + Added error checking support
# ------------------------------------------------------------------------
----------
# See url for usage:
# http://www.cyberciti.biz/tips/howto-setup-lighttpd-php-mysql-
chrooted-jail.html
# ------------------------------------------------------------------------
-----------
# Set CHROOT directory name
BASE="/home/jail"

if [ $# -eq 0 ]; then
  echo "Syntax : $0 /path/to/executable"
  echo "Example: $0 /usr/bin/php5-cgi"
  exit 1
fi

[ ! -d $BASE ] && mkdir -p $BASE || :

# iggy ld-linux* file as it is not shared one
FILES="$(ldd $1 | awk '{ print $3 }' |egrep -v ^'\(')"

echo "Copying shared files/libs to $BASE..."
for i in $FILES
do
  d="$(dirname $i)"
  [ ! -d $BASE$d ] && mkdir -p $BASE$d || :
  /bin/cp $i $BASE$d
done

# copy /lib/ld-linux* or /lib64/ld-linux* to $BASE/$sldlsubdir
# get ld-linux full file location
sldl="$(ldd $1 | grep 'ld-linux' | awk '{ print $1}')"
# now get sub-dir
sldlsubdir="$(dirname $sldl)"

if [ ! -f $BASE$sldl ];
then
  echo "Copying $sldl $BASE$sldlsubdir..."
  /bin/cp $sldl $BASE$sldlsubdir
else
  :
fi
```

Bibliografia

[1] Departament de Territori i Sostenibilitat. (2014). La qualitat de l'aire a catalunya – anuari 2014, 18.

[2] Benedict, K. (2015). Short Term Measurements and Air Quality Messaging / Regulatory

Requirements for Data, 2.

[3] CAPTOR | Computer Networking Research Group. (n.d.). Retrieved March 1, 2016, from

http://compnet.ac.upc.edu/intranet/?q=node/217

[4] European Commission. (n.d.). CAPTOR : Collective Awareness Platform for Tropospheric Ozone

Pollution Objectives.

[5] CommonSense | Computer Networking Research Group. (n.d.). Retrieved March 2, 2016, from

http://compnet.ac.upc.edu/intranet/?q=node/199

[6] CommSensum Beta. (n.d.). Retrieved March 2, 2016, from http://commsensum.pc.ac.upc.edu/

[7] Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. (n.d.). Retrieved March 2,

2016, from http://www.crummy.com/software/BeautifulSoup/bs4/doc/

[8] Documentation for sparklemotion/nokogiri. (n.d.). Retrieved March 2, 2016, from

http://www.rubydoc.info/github/sparklemotion/nokogiri

[9] lxml - Processing XML and HTML with Python. (n.d.). Retrieved March 2, 2016, from http://lxml.de/

[10] Apache Tika – Apache Tika. (n.d.). Retrieved March 2, 2016, from http://tika.apache.org/

[11] DB-Engines Ranking - popularity ranking of database management systems. (n.d.). Retrieved

March 2, 2016, from http://db-engines.com/en/ranking

[12] A Comparison Of NoSQL Database Management Systems And Models | DigitalOcean. (n.d.).

Retrieved March 2, 2016, from https://www.digitalocean.com/community/tutorials/a-comparison-of-

nosql-database-management-systems-and-models#nosql-dbmss-in-comparison-to-relational-dbmss

[13] Apache Spark TM - Lightning-Fast Cluster Computing. (n.d.). Retrieved March 2, 2016, from

http://spark.apache.org/

[14] Apache TM Hadoop®! (n.d.). Retrieved March 2, 2016, from http://hadoop.apache.org/

[15] Scrum Alliance. (n.d.). Retrieved April 1, 2016, from https://www.scrumalliance.org/

[16] MichaelPage. (2016). Estudio de remuneración 2016 - Tecnología. Retrieved from

http://www.michaelpage.es/sites/michaelpage.es/files/tecnologia2016.pdf

[17] PagePersonnel. (2016). Estudios de remuneración 2016 - Tecnología. Retrieved from

http://www.pagepersonnel.es/sites/pagepersonnel.es/files/er_tecnologia16.pdf